UCLouvain

Université catholique de Louvain

Louvain School of Management
Doctoral Program in Economics and Management

Forecasting under uncertainty:
combining and evaluating predictive models
in economics and finance

Francesco Roccazzella

Thèse présentée en vue de l'obtention
du grade de Docteur en Sciences économiques et de Gestion

Committee:
Prof. Frédéric Vrins (UCLouvain), Advisor
Prof. David Ardia (HEC Montréal)
Prof. Vincent Bodart (UCLouvain), President
Prof. Walter Distaso (Imperial College London)
Prof. Roberto Renò (ESSEC Business School)

Louvain-la-Neuve, 2023

"There are two kinds of forecasters: those who don't know, and those who don't know they don't know."

— John Kenneth Galbraith

Abstract

In recent years, the field of forecasting has experienced significant growth, thanks to advancements in statistical techniques, increased data availability, and the use of open-source predictive models. While there are methods to quantify the risk or statistical uncertainty of a predictive framework, decision-makers also face Knightian uncertainty due to incomplete knowledge of the underlying data-generating process or the likelihood of unknown external shocks.

When the decision maker is asked to form an expectation of a variable of interest and multiple forecasts are available, we show that optimal and robust combinations of forecasts offer superior forecast performance. These techniques enforce a non-negativity constraint on the weights and use penalties (or shrinkage) to penalize the divergence from a reference combination scheme. Their benefits include performing forecast selection and combination in a single step while also mitigating the risk of over-fitting by shrinking the optimal but potentially unstable solution to a sub-optimal but satisfactory one.

Combinations provide useful insights also when evaluating competing forecasting strategies. We also use optimal combinations to introduce a novel forecast encompassing test under constrained parameter space. By explicitly considering that weights are constrained in the unit simplex, the proposed test has adequate size properties and is more powerful than a competitive test that ignores such constraints. We apply these newly developed combination and evaluation techniques to tackle operational challenges in the credit industry and when evaluating monetary policy in the euro area. Model risk presents a significant challenge for both practitioners and regulators in the credit industry, as there is a risk of a mismatch between the expected and actual results of a model. Decisions concerning model risk include selecting the most appropriate modeling framework, and predictors to employ. We find that combining models that are built on security-specific characteristics and a limited number of established recovery rate determinants permits a favorable trade-off between accurate predictions and interpretability. We also stress the relevance of incorporating economic uncertainty measures and economic predictors in designing recovery rate forecast methods for corporate bonds and consumer credit.

In some circumstances, however, the decision maker is confronted with the likelihood of a *new* event that, despite impacting the target of interest, is overlooked by the available models. In this case, uncertainty arises from the unavailability of supported forecasts. We show how drawing comparisons between the new event and past experiences can aid decision-making. For instance, we design a model that describes the cross-border shock transmission mechanism in the European sovereign bond market during the Eurozone debt crisis. This helps us to evaluate the risk of a potential second crisis during the COVID pandemic.

While the proposed methods are tailored to financial and economic forecasting, this thesis offers guidance to the broader forecasting community when dealing with uncertainty.

Acknowledgments

I would like to express my gratitude to the individuals and institutions who have contributed to the successful completion of this doctoral dissertation. Their support has been indispensable throughout this journey.

First, I am grateful to my supervisor, Prof. Frédéric Vrins. I thank you for welcoming me to your research team, for having many stimulating discussions, being patient, available, and supportive. You allowed me to work under the best conditions, building a stimulating research environment, based on trust, cooperation, and fairness.

I extend my sincere appreciation to the members of my supervisory panel and thesis committee. I thank Prof. David Ardia and Prof. Vincent Bodart for your critical insights, constructive suggestions and for the time and effort you invested in reviewing my dissertation. Special thanks go to Prof. Walter Distaso and Prof. Roberto Renò for having accepted to be part of my thesis committee. Prof. Distaso, you contributed to many bright research ideas and helped me to have access to unique research opportunities. Working with you has been a real pleasure and having you as a coauthor is an honor. I had the pleasure of having Prof. Roberto Renò as the instructor of the financial mathematics course at the University of Siena. Since then, you have been my reference in *academia*. Your passion for teaching and research are truly inspiring and motivating. Therefore, having you on my thesis committee has a very special meaning for me.

I would also like to extend my sincere thanks to Prof. Bertrand Candelon, who is not a part of my dissertation committee, but whose guidance, expertise, and feedback have been decisive in shaping my research interests and completing this dissertation. You were always available when I needed help, and supportive throughout my doctoral journey and I do not have words to describe how important you have been during these last four years. I also would like to thank Prof. Christian Brownlees for your precious help, the many insightful discussions, and the hospitality you offered me at the Universitat Pompeu Fabra in Barcelona. Your critical feedback and your willingness to share your knowledge have greatly contributed to the development and refinement of my research ideas.

I wish to thank Prof. Alberto Dalmazzo, Prof. Massimo De Francesco, Prof. Carlo Pacati, and Prof. Igor Master from my former studies at the University of Siena and the University of Ljubljana for transmitting me the passion for economics and finance. You did an amazing job and I owe you a lot. A special mention goes to my previous colleagues in the research department ARC of Banka Slovenije and in particular to the director Dr. Arjana Brezigar Masten for allowing me to explore the world of central banking for the first time. You taught me many valuable lessons and I will always bring my days in Slovenia painted on my heart.

I am thankful to the faculty and staff of UCLouvain, particularly the LIDAM research institute, the LFIN and CORE research centers for providing a stimulating academic environment and the necessary resources for my research. The intellectual discussions, seminars, and workshops have played a vital role in shaping my research skills and broadening my horizons. Special thanks go to Raphaël Tursis for his valuable IT

support and to Catherine Germain, Marie Gonze, and Margaux Hubin for their priceless administrative support. I would also like to acknowledge the financial support provided by the Economic School of Louvain, the F.N.R.S., and the Fédération Wallonie-Bruxelles. Their funding has been instrumental in facilitating my research activities, including travel expenses, and access to necessary resources. Their investment in my academic pursuits has been invaluable, and I am deeply grateful for their support.

My colleagues played an important role in the completion of this doctoral thesis. A special mention goes to Angelo Luisi, Paolo Gambetti, Jonas Striaukas, and Jean-Charles Wijnandts. Together, we navigated through some *complex* situations, and I am glad to see each of us completing our respective doctoral journeys, attaining remarkable outcomes, and fulfilling our predetermined objectives. You are not only colleagues, but friends, mentors, co-authors, and the best group in which I could have ever dreamed of working. I would also like to thank to Nira, Tiziano, Fabrizio, Charles, Erika, Sonia, Camilla, Cheikh, Rim, Pavel, Roland, Farah and Liana the other members of LFIN group. I would not be writing these lines, if it was not for you.

It is difficult to list every friend that made this journey possible, from the amazing friends I have met during my studies in Siena, to the friendships I have made in my hometown. However, a special thank goes to Mario, I will always keep the memories of my trip to the University of Chicago, your warm welcome in Evanston, and our long calls in my heart. I hope you can see this from wherever you are.

In conclusion, I thank my parents, grandparents, and my brother for their unwavering support, understanding, and belief in my abilities. Thank Eddy, Sabine, Simon, and Papy Jean, for your kindness and welcoming me into your family. And Sarah, you have been the cornerstone of these last seven years from the Erasmus program in Slovenia to the completion of the PhD. Thank you for always being there for me.

This doctoral thesis is dedicated to all of you.

Contents

Chapter 1

Introduction and summary

The art and science of forecasting

The forecasting literature (Makridakis, 1986) has evolved over the last decades. The study of predictive methods in the areas of demography, energy, transportation, meteorology, economics, and finance is burgeoning, boosted by the recent development of statistics and ease of access to data, computing power, and the availability of open-source libraries of predictive models. Scientists and professional forecasters, however, still build predictive models with a particular application in mind: predicting astronomical or meteorological phenomena is different from predicting the dynamics of the economic and business cycles. Overlooking the differences between applications and ignoring the specific modeling assumptions on which predictive frameworks are built becomes a cardinal sin. Decision-makers are aware of this issue and they factor it in when forming their expectations.

The science of forecasting provides the tools to quantify the *risk* or *statistical uncertainty* of *a* predictive framework by looking at the probability distribution of its realized forecast errors. This requires knowing the probability distribution of the variable of interest. However, if the probability distribution is unknown, decision-makers are confronted with *Knightian (or true) uncertainty*.

Following Watkins and Knight (1922), we define Knightian uncertainty as the fundamental degree of ignorance and the essential unpredictability of future events. Knightian uncertainty derives from the incomplete discovery of the underlying data-generating process (DGP) of the variable of interest. Incomplete learning may result from i) partial identification of the determinants or predictive variables driving the dynamics of the target variable, ii) partial understanding of the functional form linking the target variable and its determinants, iii) data limitations, and, iv) time-series applications, the presence of structural breaks in the dynamics of the target. Alternatively, incomplete learning may derive from the likelihood (risk) of a new external shock occurring and affecting the target of interest.[1] In this case, the forecaster defines *scenarios*, estimating *what* would be the effect on the target of interest *if* the shock occurs. As a result of incomplete learning, decision-makers are often confronted with two exact opposite scenarios: the availability of many or no (or unsupported) forecasts to guide their final decision. Thus, despite the advances of predictive modeling, forecasters' experience and subjective judgment remain at the core of the discipline, hence the expression *the art of*

[1] For example, we can be interested in whether the weakening of the financial conditions following the pandemic could trigger a second sovereign debt crisis in the euro area.

forecasting.

Machine learning and big data have guaranteed access to *many* predictions for a variable of interest and have significantly contributed to a generalized improvement of predictive methods in terms of quantifiable risk. However, they have also raised questions on how to identify the appropriate predictive strategy for the application of interest among the *universe* of available models and the *galaxy* of predictive variables. The wording *universe* and *galaxy* is intentional. Focusing only on supervised machine learning algorithms for classification tasks, Caruana and Niculescu-Mizil (2006) list 10 broad categories of learning methods: SVMs, neural nets, logistic regression, naive Bayes, memory-based learning, random forests, decision trees, bagged trees, boosted trees, and boosted stumps. However, each method has potentially many configurations, which results in 2,000 models under analysis. The number of predictions at hand would grow exponentially if we also include model combinations and variable-selection techniques as distinct model specifications.

Such volume of information can certainly help an individual to make an informed decision and, in particular, when predictions from different forecasting strategies suggest similar outcomes. Nevertheless, human beings must make the final decision also when the available forecasts suggest contradictory outcomes. In this case, uncertainty raises and the human ability to identify the relevant piece of information can be undermined. The cognitive capability of the mind and the time available to make the decision limit individuals' ability to make an *optimal* decision. Because of these limits, rationality is bounded when individuals make decisions, and consequently, decision-makers seek a *satisfactory* solution that relies on heuristics and previous experience rather than the optimal solution that would instead result from the full cost-benefit analysis (Simon, 1957). For these reasons, practitioners often disregard sophisticated forecasting techniques and favor more tractable frameworks or strategies to which they have become accustomed. However, academic research is providing growing evidence of the potential of machine learning in these regards.[2] More work needs to be done to design methods that are not perceived as *black boxes* and that take model uncertainty explicitly into account.

Building on finance's portfolio theory and the literature on forecast combination and forecast encompassing, this doctoral thesis aims to define a unified framework for the design and evaluation of predictive strategies in economics and finance when the forecaster is confronted with risk and Knightian uncertainty and multiple forecasts are available. Indeed, forecasts combination have received increasing attention in the forecasting community (Makridakis, Spiliotis, and Assimakopoulos, 2020; Atiya, 2020; Petropoulos, Apiletti, Assimakopoulos, Babai, Barrow, Taieb, Bergmeir, Bessa, Bijak, Boylan, Browell, Carnevale, Castle, Cirillo, Clements, Cordeiro, Oliveira, Baets, Dokumentov, Ellison, Fiszeder, Franses, Frazier, Gilliland, Gönül, Goodwin, Grossi, Grushka-Cockayne, Guidolin, Guidolin, Gunter, Guo, Guseo, Harvey, Hendry, Hollyman, Januschowski, Jeon, Jose, Kang, Koehler, Kolassa, Kourentzes, Leva, Li, Litsiou, Makridakis, Martin, Martinez,

[2]For example, in credit risk modeling, nonlinear techniques (Loterman, Brown, Martens, Mues, and Baesens, 2012; Yao, Crook, and Andreeva, 2015) and ensemble methods (Hartmann-Wendels, Miller, and Töws, 2014; Bellotti, Brigo, Gambetti, and Vrins, 2021) should be preferred to the traditional parametric regressions used in earlier studies on recovery rate determinants.

Meeran, Modis, Nikolopoulos, Önkal, Paccagnini, Pan, 2022). Combining weights can be estimated from realized (in-sample) forecast errors by minimizing a predefined loss function, thereby minimizing statistical uncertainty. However, the forecaster also faces Knightian uncertainty as the forecast errors may be subject to structural breaks in their past and future dynamics.[3] Shrinking the optimal combination towards a fixed weighting scheme and restricting the weights to lay in the unit simplex, i.e., constraining the weights to be nonnegative and sum to one, mitigate the risk of *overfitting*[4] and improve predictive performance. We empirically show how these modern portfolio optimization techniques outperform those considered in previous forecasting applications in extensive simulation studies and empirical exercises in economics (Roccazzella, Gambetti, and Vrins, 2022b).

Given a loss function and by examining the probability distribution of the realized forecast errors, i.e., when focusing on statistical uncertainty, the combining weights also offer useful insights when the forecaster is interested in evaluating competing forecasting strategies. Indeed, forecast combination and the forecast encompassing literature are closely interconnected, and Diebold (1989) made the first attempt to formally reconcile these two streams of literature. On the one hand, combining weights reflect forecasts' marginal contribution to the reduction of the combination's loss. On the other hand, forecast encompassing aims to design evaluation strategies to detect whether a forecast combination incorporates more information than the individual forecasts (Clements and Hendry, 1998). Therefore, when looking at forecast encompassing and forecasts combination from a joint perspective, combining weights can be naturally interpreted as the marginal contribution of each forecaster in terms of independent information. Following this direction, we use optimal combinations of forecasts to introduce a novel forecast encompassing test under constrained parameter space. By explicitly considering that weights are constrained in the unit simplex, the proposed test proposes adequate size properties and it is more powerful than a competitive test that ignores such constraints. Moreover, the testing procedure can be extended to frameworks in which the number of competing forecasts outclasses the number of available realizations of the target variable (Roccazzella and Candelon, 2022).

We have previously assumed that the forecaster is confronted with multiple forecasts. The multiplicity and heterogeneity of the forecasts themselves are the sources of uncertainty: the forecaster has to process the available information and form his/her expectation. However, Knightian uncertainty is also associated with situations that involve new or unique events where there is no historical data or reliable forecasting models to draw upon. This setting contrasts with the literature on parameter instability and how to make optimal use of past data (and forecast combinations) to forecast out-of-sample (Pesaran, Pettenuzzo, and Timmermann, 2006; Pesaran and Pick, 2011; Pesaran, Pick, and Pranovich, 2013) even when taking into account the possibility that breaks have occurred in the past may re-occur in the future (see for example Maheu and Gordon, 2008). Under these circumstances, a general forecasting method can hardly be defined and the forecaster should instead design *ad hoc* strategy to investigate the mechanism through which the

[3]The breaks in the data generating process may also occur in cross-sectional studies with heterogeneous in-sample and out-of-sample data sets.

[4]Overfitting is the estimation of a model that corresponds too closely to a specific sample, consequently undermining its ability to generalize its predictions to future or novel observations in a reliable manner.

new event influences the target of interest. In this doctoral thesis, we highlight how via the identification of the analogies of the new phenomenon with previous experiences, the forecaster may validate the resulting *scenario*, guiding the decision-making process. For example, by building a model that *ex-ante* described the inter-country shock transmission mechanism in the euro area sovereign bond market following the outburst of the euro area debt crisis in 2010, we could monitor the risk of a second European Sovereign Debt crisis during (and following) the COVID pandemic. In Candelon, Luisi, and Roccazzella (2022), we introduce a novel methodology that combines Factor and Global Vector Autoregressive models to show that fragmentation risk well preceded the sovereign debt crisis outburst and, by analyzing the recent period, to document a rise in fragmentation risk in the euro area during the COVID pandemic.

Structure of the Thesis

Figure 1.1 describes the structure of this doctoral thesis, highlighting the linkages between the five main chapters. On the left side of the dashed separating line, we deal with issues deriving from the existence of multiple forecasts.

Chapter 2 reviews the literature on forecast combination and we introduce various methods that combine forecasts using constrained optimization with a penalty. The proposed framework performs forecast selection and combination in one step, allowing for potentially sparse combining schemes. We provide the analytical expression of the optimal penalty strength when penalizing with the L2 divergence from the equally-weighted scheme and investigate the impact of the divergence function, the reference scheme, and the non-negativity constraint on the predictive performance via an extensive simulation study and two empirical applications.

Chapter 3 discusses the use of model combinations to identify the most effective algorithms and predictive variables for corporate bond recovery rate prediction. We find that the most promising approach involved combining models trained on security-specific characteristics and a limited number of well-established recovery rate determinants, including measures of economic uncertainty.

Chapter 4 examines the determinants of loss given default (LGD) for consumer credit and challenges previous research that suggested only contract-specific variables were relevant. The study used a much larger dataset than previous studies, including more than 6 million Italian consumer debts from 2007 to 2019, and over 40 predictors. We find that regional social and welfare indicators and credit cycle variables are significant contributors to LGD forecasting performance, enhancing it by up to 15 p.p. in terms of R^2. We also find that nonlinear combination schemes based on artificial neural networks are the most effective approach for predicting recovery rates for both individual and portfolios of consumer credits.

Chapter 5 connects optimal combining weights to the field of forecast evaluation. Specifically, we use optimal combinations of forecasts to introduce a novel forecast encompassing test to evaluate time-series and institutional inflation projections in

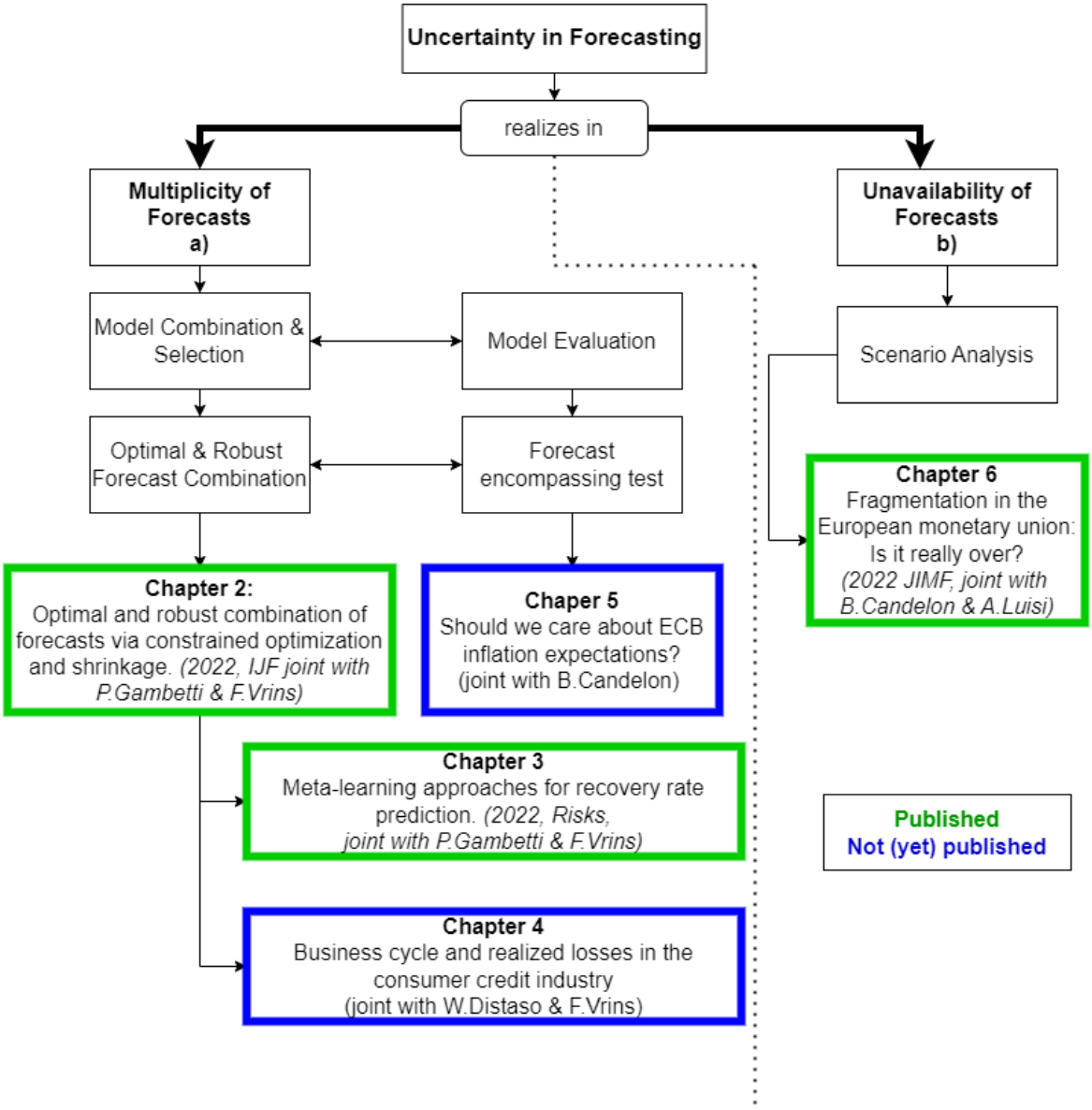


Figure 1.1: Thesis structure and linkages with scientific articles.

the euro area. In a forecast encompassing test, two (or more) competing forecasts are compared by testing the null hypothesis that one is not worse than any linear combination of the forecasts from all the considered models. Hence, combination weights reveal which forecasts are the most informative. We find that although the ECB is the most informative forecaster on average, it does not encompass its competitors and its weight varies over time. Macro-financial conditions and monetary policy actions explain this variability. The greater the uncertainty surrounding inflation and the difference between the current and the 2% inflation target, the less informative ECB's forecasts are. The more contractionary the monetary policy, the more informative they are. We find that ECB's declining weight and its relation with its determinants raise a warning flag: the potential loss of informativeness damages the central bank leading role in anchoring inflation expectations and questions whether the goal of preserving financial stability is compatible with the inflation targeting objective.

On the right side of the dashed separating line, we deal with uncertainty deriving from the unavailability of forecasts. Specifically, we answer the *what if* question, and explain how through the identification of the analogies of the new phenomenon with previous experiences, the forecaster may validate the resulting *scenario*, guiding the decision-making process.

Chapter 6 applies this idea to analyze fragmentation risk in the European Monetary Union, investigating the shock transmission mechanisms in the sovereign bond market. To achieve this goal, it builds a new methodology for modeling interactions, reconciling Factor and Global Vector Autoregressive models. This framework helps to disentangle interdependence from contagion and flight-to-quality effects, therefore assessing more precisely the degree of fragmentation. When looking at sovereign bond debt markets in the euro area (EA), fragmentation risk is the risk that sovereign bonds spreads across EA countries display a heterogeneous reaction to a common shock or a *foreign* shock, i.e., a shock hitting a specific country (Candelon, Luisi, and Roccazzella, 2022). It turns out that fragmentation risk was already present and sizable in the pre-European sovereign debt crisis period, but has been sharply mitigated afterward. The COVID pandemic has put it back at the forefront of the scene, questioning the European integration process and calling for adequate policy measures.

List of scientific contributions and published material

The following articles, co-authored during the writing of this thesis, have influenced this work.

- Roccazzella, F., P. Gambetti, and F. Vrins (2022b). Optimal and robust combination of forecasts via constrained optimization and shrinkage. *International Journal of Forecasting 38* (1), 97-116.[5]
- Gambetti, P., F. Roccazzella, and F. Vrins (2022). Meta-learning approaches for recovery rate prediction. *Risks 10* (6).
- Roccazzella, F. and Candelon, B (2022). Should we care about ECB inflation expectations? *LIDAM Discussion Paper LFIN*, 2022/04.
- Candelon, B., A. Luisi, and F. Roccazzella (2022). Fragmentation in the European monetary union: Is it really over? *Journal of International Money and Finance 122*, 102545.

[5]For the correction of Proposition 1, please see: Roccazzella, Gambetti, and Vrins (2022a).

Chapter 2

Optimal and robust combination of forecasts

with Paolo Gambetti and Frédéric Vrins

Abstract

We introduce various methods that combine forecasts using constrained optimization with penalty. A non-negativity constraint is imposed on the weights, and several penalties are considered, taking the form of a divergence from a reference combination scheme. In contrast with most of the existing approaches, our framework performs forecasts selection and combination in one step, allowing for *potentially* sparse combining schemes. Moreover, by exploiting the analogy between forecasts combination and portfolio optimization, we provide the analytical expression of the optimal penalty strength when penalizing with the L2-divergence from the equally-weighted scheme. An extensive simulation study and two empirical applications allow us to investigate the impact of the divergence function, the reference scheme and the non-negativity constraint on the predictive performance. Our results suggest that the proposed models outperform those considered in previous studies.

Introduction

Often, m different models can be used to produce a set of forecasts $\widehat{\boldsymbol{y}} = (\widehat{y}_1, \dots, \widehat{y}_m)^T$ of a given random variable y. In such cases, two strategies can be considered to obtain an aggregated forecast $\widehat{y}$: the best prediction can be chosen (i.e., $\widehat{y} = \widehat{y}_{i^\star}$, where model $i^\star \in \{1, \dots, m\}$ is selected according to some criterion), or the forecasts can be combined into a single prediction (i.e., $\widehat{y} = \Phi(\widehat{\boldsymbol{y}})$, for some function $\Phi : \mathbb{R}^m \to \mathbb{R}$).[1]

The first strategy usually provides disappointing results, and it is now widely recognized that combining models is preferable (Granger and Ramanathan, 1984; Atiya, 2020; Taylor, 2020). Most combination strategies include two stages. The first stage consists of removing poorly performing models via trimming based on rules of thumb (e.g., retaining a fixed percentage of models as in Aiolfi and Favero, 2005), formal tests (Samuels and Sekkel, 2017; Amendola, Braione, Candila, and Storti, 2020), measures of interest (Granger and Jeon, 2004), or machine learning methods (Diebold and Shin, 2019) so that only k models out of m candidates are kept. Then, the selected candidates are combined using a combination scheme, usually in a linear way (i.e., $\Phi(\widehat{\boldsymbol{y}}) = \boldsymbol{w}^T\widehat{\boldsymbol{y}}$ where $\boldsymbol{w}$

[1] Throughout the chapter, we use the terms "forecast", "prediction" and "model" as synonyms. We denote matrices as bold capital letters, while vectors and scalars are represented by bold lowercase letters and lowercase letters, respectively.

is a m-by-1 vector of combining weights).

When the individual forecasts are unbiased, the standard approach consists of minimizing a loss function featuring the aggregated forecast error under the constraint that the weights sum to 1.[2] For instance, the minimum-variance combination scheme $\boldsymbol{w}^*$ is obtained by minimizing $\mathbb{E}\|y - \boldsymbol{w}^T\widehat{\boldsymbol{y}}\|_2^2$ under the constraint $\boldsymbol{w}^T\mathbf{1} = 1$ or, equivalently (Granger and Ramanathan, 1984):

$$\boldsymbol{w}^* := \underset{\boldsymbol{w} \in \mathscr{W}}{\arg\min}\, \boldsymbol{w}^T\boldsymbol{S}\boldsymbol{w} \quad \text{where} \quad \mathscr{W} := \{\boldsymbol{w} \in \mathbb{R}^m \mid \mathbf{1}^T\boldsymbol{w} = 1\}\,, \tag{2.1}$$

where $\boldsymbol{S}$ is the population covariance matrix of models' prediction error, $v_i := y - \widehat{y}_i$. The solution to (2.1) is

$$\boldsymbol{w}^* = \Psi(\boldsymbol{S}) \quad \text{where} \quad \Psi(\boldsymbol{M}) := \left[\mathbf{1}^T\boldsymbol{M}^{-1}\mathbf{1}\right]^{-1}\boldsymbol{M}^{-1}\mathbf{1}\,.$$

In practice, $\boldsymbol{S}$ is typically unknown and must be estimated from the data. Clearly, given a sample of n observations, the quality of the optimal scheme depends to a large extent on the estimator of $\boldsymbol{S}$. One possibility is to consider the sample counterpart of $\boldsymbol{S}$, i.e., $\boldsymbol{\Sigma} := \frac{1}{n}\boldsymbol{V}^T\boldsymbol{V}$, where $\boldsymbol{V}$ is the $n \times m$ matrix of realized prediction errors. However, despite being consistent, $\boldsymbol{\Sigma}$ breaks down in high dimensional frameworks ($n \ll m$) and can be imprecise due to sample size limitations or the presence of considerable background noise. Hence, the estimated $\boldsymbol{w}^*$ will be unstable and the corresponding predicting strategy will perform poorly out of sample (Smith and Wallis, 2009).

Previous literature attempted to tackle this issue using two distinct approaches. The first approach is to rely on regularization (or shrinkage) techniques such as constrained optimization with penalty (COP), with shrinkage towards equal weights (Diebold and Pauly, 1990; Diebold and Shin, 2019).[3] The second route is to impose some constraints on the combining weights, such as restricting them to be non-negative (Conflitti, Mol, and Giannone, 2015). In this respect, it is worth noting that this field shares many similarities with that of portfolio optimization in finance. The popular investment strategy called the *minimum-variance portfolio* (Markowitz, 1952) precisely corresponds to (2.1) and its performance suffers from the same shortcomings regarding the estimator $\boldsymbol{\Sigma}$ (Michaud, 1989; DeMiguel, Garlappi, and Uppal, 2007).[4] Shrinkage techniques on the weights (DeMiguel, Garlappi, Nogales, and Uppal, 2009) or no-short sales constraints corresponding to the non-negativity restriction (Jagannathan and Ma, 2003) are similarly applied in that domain. Interestingly, however, the portfolio optimization literature also brings further solutions. For instance, it has been proposed to *combine* the weights resulting from advanced (i.e., asymptotically unbiased but potentially high-variance) and naive (i.e., biased but low-variance, such as the equally weighted) schemes (Tu and Zhou,

[2] We have assumed individual forecasts to be unbiased throughout the chapter, but this assumption can be easily relaxed. If we reformulate the optimal forecast combination problem in the form of a linear regression, the bias can be easily corrected by adding the intercept to the model as discussed in Timmermann (2006).

[3] Diebold and Shin (2019) extend the standard COP method using a two step procedure: having selected a subset of forecasts using LASSO, the weights of the selected forecasts are estimated by penalizing the L1 or L2 norm of the difference between the estimated and equal weights.

[4] In that context, the vector of weights $\boldsymbol{w}$ represents the proportion of wealth assigned to each asset, $\boldsymbol{S}$ represents the population asset returns' covariance matrix and $\boldsymbol{\Sigma}$ is its sample estimator.

2011). Another proposed method is instead to alleviate the noise impacting $\boldsymbol{\Sigma}$ by using the *shrinkage of the covariance matrix*, a convex combination of the sample covariance matrix $\boldsymbol{\Sigma}$ and of a highly structured covariance matrix $\overline{\boldsymbol{\Sigma}}$ (Ledoit and Wolf, 2004a).[5] These methods have never been explored in a forecasts combination context. In particular, existing combination schemes still strongly rely on the noisy estimator $\boldsymbol{\Sigma}$ of $\boldsymbol{S}$ (Claeskens, Magnus, Vasnev, and Wang, 2016).

More generally, the literature on forecasts combination still suffers from several limitations. First, (i) by featuring a model selection step, existing methods implicitly assume the best forecasting strategy to be a *sparse* combination of the candidate forecasts and/or they still rely on the *a priori* identification of the forecasts to combine. The decision regarding how to trim the set of model is generally arbitrary and can deteriorate the quality of the combination (Winkler and Makridakis, 1983)[6]. Second, (ii) COP schemes and the non-negativity constraint have been explored independently so far; it is unclear under what conditions combing the two might be beneficial. Third, (iii) existing COP methods feature L1 or L2 penalty to shrink the solution towards a reference scheme but other divergence measures (possibly tailored to deal with non-negative weights) remain unexplored. Finally, (iv) methods that shrink the solution towards a reference scheme often rely on cross-validation to estimate the penalty strength. However, cross-validation can be harmful in small sample, and its properties are not trivial when there is auto-correlation in forecast errors.

In this chapter, we address the above gaps as follows. Regarding (i), we design a robust single-step framework that allows to obtain *potentially sparse* combining weights. This contrasts with the standard two-step procedures which *a priori* assume the combination of forecasts to be sparse (Diebold and Shin, 2019; Samuels and Sekkel, 2017; Aiolfi and Favero, 2005). Regarding (ii), we propose for the first time to combine the non-negativity constraint with COP methods leading to COP$^+$ schemes. This restriction is shown to blend forecast selection and weights estimation in a single-step and helps mitigating estimation uncertainty (in line with the observations made in the literature on portfolio theory). We also present and discuss the scenarios where dropping the non-negativity restriction can be beneficial. We find that this occurs only under specific circumstances: when a candidate's prediction errors are strongly positively correlated to those of the other forecasts and when, at the same time, its predictive ability is also significantly inferior. Moreover, we contribute to (iii) by introducing new divergence measures in COP (namely, elastic net and cross-entropy) and study their performance. We find promising results for cross-entropy (only valid when working under the non-negativity constraint), which consistently outperforms the standard L1 penalization. Eventually, (iv) we use

[5]There are deep connections between these approaches, especially when minimum variance is the objective function. For instance, $\overline{\boldsymbol{\Sigma}} = \boldsymbol{I}$ amounts to penalize the Euclidean distance between the weights and the equally-weighted scheme (DeMiguel, Garlappi, Nogales, and Uppal, 2009). Similarly, setting an upper limit on the L2-norm corresponds to add a penalty to prevent the weights to be too far from the equally-weighted scheme (DeMiguel, Garlappi, Nogales, and Uppal, 2009). Alternatively, imposing a constraint on the L1-norm on the weights amounts to add a no-shortsales constraint (DeMiguel, Garlappi, Nogales, and Uppal, 2009). This latter also corresponds to reducing the large elements of the sample covariance matrix (Jagannathan and Ma, 2003).

[6]For instance, forecasts tend to be highly correlated, and statistical procedures such as LASSO often fail to correctly select the best subset of forecasts (Zhao and Yu, 2006)

the correspondence between the shrinkage estimators of the sample covariance matrix (Ledoit and Wolf, 2004a) and the L2-norm penalty (DeMiguel, Garlappi, Nogales, and Uppal, 2009) to obtain an analytical expression for the optimal penalty strength, thereby making the economy of a potentially harmful cross-validation step in this case. In an extensive simulation study and in two empirical applications, we find that these "explicit COP models" outperform cross-validation based COP methods, the simple average forecast, trimming methods and other standard benchmarks from previous studies. We document the flexibility and robustness of the single-step combinations with explicit penalty strength, methods called *optimal and robust combinations of forecasts* hereafter.

The remainder of this chapter is structured as follows. Section 2.1 describes the proposed optimal and robust combination of forecasts. Section 2.2 includes the simulation study. Section 2.3 is dedicated to the empirical applications. Section 2.4 concludes the chapter.

2.1 The one-step optimal and robust combination of forecasts

We now study the features of the proposed optimal and robust combination of forecasts.

2.1.1 The optimal combination of forecasts with non-negative weights

The proposed methods aim to blend forecast selection and weights estimation in a single step. Let us briefly explain how and why constraining the weights to be non-negative and to sum to 1 implicitly selects which forecasts to combine. Intuitively, if a candidate forecast j has a higher covariance with other strategies than other candidates, the optimization procedure will reduce w_j accordingly and eventually set it to 0 if the marginal contribution of $\hat{y}_j$ to the variance of the aggregate prediction error is always larger than the other forecasts' marginal contributions. Furthermore, imposing non-negativity and portfolio constraints also reduces the estimation error by inducing a shrinkage-like effect on the sample covariance matrix (Jagannathan and Ma, 2003).

We consider the following constrained minimization problem (the covariance matrix is always assumed to be positive definite in population in the sequel):

$$\boldsymbol{w}^* := \underset{\boldsymbol{w} \in \mathscr{W}^+}{\arg\min}\ \boldsymbol{w}^T \boldsymbol{\Sigma} \boldsymbol{w} \quad \text{where} \quad \mathscr{W}^+ := \{\boldsymbol{w} \in \mathscr{W} \mid \min \boldsymbol{w} \geq 0\}\,. \tag{2.2}$$

This expression is similar to (2.1) except that the optimization set is now $\mathscr{W}^+$ (the set of non-negative weights summing to one) and that the (unknown) covariance matrix $\boldsymbol{S}$ has been replaced by its sample counterpart, $\boldsymbol{\Sigma}$. The optimal solution is known analytically and can be retrieved by imposing the Karush-Kuhn-Tucker (KKT) conditions (Proposition 1 of Jagannathan and Ma, 2003):

Proposition 2.1.1. *The solution to the optimization problem* (2.2) *is given by*

$$\boldsymbol{w}^* := \Psi^+(\boldsymbol{\Sigma}) \quad \textit{where} \quad \Psi^+(\boldsymbol{M}) := \frac{\boldsymbol{M}_+^{-1}\mathbf{1}}{\mathbf{1}^T \boldsymbol{M}_+^{-1}\mathbf{1}} \tag{2.3}$$

where $\mathbf{1}$ *is a conformable vector of ones and* $\boldsymbol{M}_+ := \boldsymbol{M} - \left(\mathbf{1}\boldsymbol{\mu}^T + \boldsymbol{\mu}\mathbf{1}^T\right)$ *with* $\boldsymbol{\mu}$ *being the* m*-dimensional vector of Lagrange multipliers corresponding to the nonnegativity constraints in the KKT conditions.*

Conflitti, Mol, and Giannone (2015) notice that Proposition 2.1.1 is a special case of LASSO regression since the non-negativity restriction and the sum to one constraint imply that the weight vector belongs to the unit simplex, and therefore weights have a unitary L1-norm.

2.1.2 Constrained optimization with penalty (COP) on the weighting scheme

We introduce the COP combination scheme as the solution to an optimization problem on $\mathscr{W}^+$. Similarly to (2.2), we directly consider the sample covariance matrix of prediction errors in the objective function. The minimum-variance objective function is associated with a penalty term, whose purpose is to shrink the solution towards a predetermined reference scheme $\overline{\boldsymbol{w}} \in \mathscr{W}^+$ using some divergence measure. To study the impact of the non-negativity constraint, we define COP$^+$ and COP as follows.

Definition 2.1.1 (Constrained optimization with penalty - COP)**.** *The COP weighting scheme with non-negativity constraint, COP*$^+$*, is defined as*

$$\boldsymbol{w}_\delta^* := \underset{\boldsymbol{w} \in \mathscr{W}^+}{\arg\min} \ \boldsymbol{w}^T \boldsymbol{\Sigma} \boldsymbol{w} + \delta d\left(\boldsymbol{w}, \overline{\boldsymbol{w}}\right) \ , \tag{2.4}$$

where $\delta \geq 0$ *is the* penalty strength *and* $d\left(\cdot,\cdot\right) := \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}^+$, $\left(\boldsymbol{u}, \boldsymbol{v}\right) \mapsto d\left(\boldsymbol{u}, \boldsymbol{v}\right)$ *is a* divergence measure *between two vectors. COP takes the same form as* (2.4) *but where* $\mathscr{W}^+$ *is replaced by* $\mathscr{W}$*.*

The coefficient δ controls the impact of the penalty, whereas $d\left(\boldsymbol{w}, \overline{\boldsymbol{w}}\right)$ quantifies the "distance" between $\boldsymbol{w}$ and the reference scheme $\overline{\boldsymbol{w}}$. The divergences that will be considered in our simulation study and in the empirical applications are:

- elastic net (EN): $d\left(\mathbf{u}, \mathbf{v}\right) = a \left\|\mathbf{u} - \mathbf{v}\right\|_1 + (1-a) \left\|\mathbf{u} - \mathbf{v}\right\|_2^2$ where $a \in [0,1]$,
- cross-entropy (CE): $d\left(\mathbf{u}, \mathbf{v}\right) = \sum_{j=1}^m u_j \ln\left(\frac{u_j}{v_j}\right)$ where $\boldsymbol{u}, \boldsymbol{v} \in \mathscr{W}^+$.

Notice that the L1 and L2 distances are special cases of EN with $a = 1$ and $a = 0$, respectively, and do not require to impose non-negativity, in contrasts with CE.

2.1.3 The shrinkage direction of combining weights

Potentially, any vector satisfying $\overline{\boldsymbol{w}} \in \mathscr{W}^+$ can be taken as $\overline{\boldsymbol{w}}$ in (2.4); for example, $\overline{\boldsymbol{w}}$ can represent a combination strategy that the forecaster believes to be suitable for a particular predictive problem or can express some prior belief about the exclusion of a model from the set of candidate forecasts. Nevertheless, in practice, we often prefer a fixed reference point to alternatives that require the estimation of individual losses, as the latter can be subject to estimation error. In particular, the simple average $\overline{\boldsymbol{w}}^E := \mathbf{1}/m$ seems a natural reference scheme. Assigning equal weights often performs quite well for both finite samples (Smith and Wallis, 2009) and the population (Aruoba, Diebold, Nalewaik, Schorfheide, and Song, 2012). Moreover, the equally-weighted scheme is rather intuitive: shrinking the weights towards $\overline{\boldsymbol{w}}^E$ amounts to penalize for the L2-norm of the weights,

since the constraint $\left\|\boldsymbol{w}-\overline{\boldsymbol{w}}^{E}\right\|_2^2$ coincides with $\|\boldsymbol{w}\|_2^2$ on $\mathscr{W}$. Finally, this choice can also offer some computational advantages, as it often simplifies the optimization problem. For example, the COP estimators of the combining weights are known in closed-form (see A2.1 for a simple proof):

Proposition 2.1.2. *In the special case where $\overline{\boldsymbol{w}}=\overline{\boldsymbol{w}}^{E}$ and $d(\boldsymbol{u},\boldsymbol{v})=\|\boldsymbol{u}-\boldsymbol{v}\|_2^2$, the COP^{+} and COP weighting schemes in Def. 2.1.1 are respectively given by :*

$$\boldsymbol{w}_{\delta}^{*}=\Psi^{+}(\boldsymbol{\Sigma}_{\delta}) \quad or \quad \Psi(\boldsymbol{\Sigma}_{\delta}) \qquad where \quad \boldsymbol{\Sigma}_{\delta}:=\boldsymbol{\Sigma}+\delta\boldsymbol{I}\,, \tag{2.5}$$

with $\boldsymbol{I}$ the m-by-m identity matrix.

Nevertheless, shrinking the combining weights towards equality can significantly damage the predictive performance if the individual prediction error variances truly differ among candidate forecasts. In this case, considering the inverse-loss weighted average as reference, i.e., $\overline{\boldsymbol{w}}=\overline{\boldsymbol{w}}^{IL}$ where $\overline{w}_i^{IL}:=[\sigma_i^2\sum_{i=1}^{m}1/\sigma_i^2]^{-1}$, $\sigma_i^2:=\Sigma_{i,i}$, is a more appealing alternative to equal weights. However, the $\overline{\boldsymbol{w}}^{IL}$ is data-driven and, therefore, is subjected to estimation error. There is no clear way to balance misspecification and estimation error in the reference scheme. Hence, in the empirical section, we test whether restricting ourselves to equal weights offers a significant disadvantage compared to the inverse-loss weighted average or whether the misspecification introduced in $\overline{\boldsymbol{w}}^{IL}$ is balanced by an increased sensitivity to estimation error in the latter.

2.1.4 Robust combination methods and the shrinkage estimator of the covariance matrix

In the previous section, we derived a forecast combination scheme that makes use of constrained optimization with penalty to estimate the optimal set of weights, but other methods can be adopted. In particular, we can consider shrinkage estimators of $\boldsymbol{\Sigma}$. This approach has been developed in a series of papers by Ledoit and Wolf (2003, 2004a,b, 2012, 2017), who introduced linear and nonlinear shrinkage techniques to reduce the instability of the estimated covariance matrix in a portfolio optimization context. As recalled in the introduction, it consists of combining the sample covariance matrix (which can be easily computed and is asymptotically unbiased but prone to estimation error) with a highly structured estimator (that contains relatively little estimation error but tends to be misspecified and biased). Transposing this technique to our forecast combination problem leads to the following combination scheme:

Definition 2.1.2 (Constrained optimization with shrinkage (COS) of the covariance matrix)**.** *The COS weighting scheme with non-negativity constraint, COS^{+}, is defined as*

$$\boldsymbol{w}_{\lambda}^{*}:=\underset{\boldsymbol{w}\,\in\,\mathscr{W}^{+}}{\arg\min}\ \boldsymbol{w}^{T}\boldsymbol{\Sigma}_{\lambda}\boldsymbol{w} \qquad where \qquad \boldsymbol{\Sigma}_{\lambda}:=(1-\lambda)\,\boldsymbol{\Sigma}+\lambda\overline{\boldsymbol{\Sigma}}, \tag{2.6}$$

$\lambda\in[0,1]$ *is the* shrinkage intensity*, and $\overline{\boldsymbol{\Sigma}}$ is the* reference covariance matrix*. COS takes the same form as* (2.6) *but where $\mathscr{W}^{+}$ is replaced by $\mathscr{W}$.*

This problem takes the same form as (2.2), hence, the solution to COS^{+} and COS are simply $\boldsymbol{w}_{\lambda}^{*}=\Psi^{+}(\boldsymbol{\Sigma}_{\lambda})$ and $\Psi(\boldsymbol{\Sigma}_{\lambda})$, respectively.

As explained above, shrinkage towards equal weights is natural and standard. Moreover, Timmermann (2006) shows that this combining scheme is optimal when the covariance matrix of the prediction errors satisfies $\Sigma_{i,i} = \sigma^2$ and $\Sigma_{i,j\neq i} = \rho\sigma^2$. Thus, shrinking towards equal weights can indirectly be achieved by shrinking the sample covariance matrix towards a special reference matrix, namely, the covariance matrix that would be implied *if equal weights indeed represented the optimal combination strategy*. Without loss of generality (see A2.2),

$$\overline{\boldsymbol{\Sigma}}^E := \overline{\sigma}^2 \boldsymbol{I}, \quad \text{where} \quad \overline{\sigma}^2 := \frac{1}{m}\sum_{i=1}^{m} \sigma_i^2 \ . \tag{2.7}$$

We can extend this framework to consider other reference matrices $\overline{\boldsymbol{\Sigma}}$ that are closely related to alternative reference weighting schemes. For example, we can indirectly obtain the inverse-loss weighted average as reference weights by shrinking the sample covariance matrix towards a specific reference $\overline{\boldsymbol{\Sigma}}^{IL}$. This requires choosing

$$\overline{\boldsymbol{\Sigma}}^{IL}_{i,i} := \sigma_i^2 \quad \text{and} \quad \overline{\boldsymbol{\Sigma}}^{IL}_{i,j} := 0, \quad \forall i \neq j \ . \tag{2.8}$$

2.1.5 A closer look at COP and COS methods.

Despite following different approaches, the optimal weighting schemes in COP and COS are closely related. In particular, in the field of portfolio optimization, DeMiguel, Garlappi, Nogales, and Uppal (2009) show that there is a correspondence between COS and the portfolio that solves the minimum-variance problem, subject to the additional constraint that the norm of the portfolio-weights vector is smaller than a given threshold. A similar relation holds in the context of constrained optimization with penalty.

Proposition 2.1.3. *The COS model* (2.6) *with reference covariance matrix* $\overline{\boldsymbol{\Sigma}} = \overline{\boldsymbol{\Sigma}}^E$ *given in* (2.7) *corresponds to the COP model* (2.4) *with* $\overline{\boldsymbol{w}} = \overline{\boldsymbol{w}}^E$, $d(\boldsymbol{u}, \boldsymbol{v}) = \|\boldsymbol{u} - \boldsymbol{v}\|_2^2$ *and* $\delta = \frac{\lambda}{1-\lambda}\overline{\sigma}^2$.

The proof of Proposition 2.1.3 is trivial. Provided that $\lambda > 0$[7], $\boldsymbol{\Sigma}_\lambda$ can be rescaled by $(1-\lambda)$ to obtain $\widetilde{\boldsymbol{\Sigma}}_\lambda := \dfrac{\boldsymbol{\Sigma}_\lambda}{1-\lambda} = \boldsymbol{\Sigma} + \dfrac{\lambda}{1-\lambda}\overline{\sigma}^2\boldsymbol{I}$. Considering $\widetilde{\boldsymbol{\Sigma}}_\lambda$ instead of $\boldsymbol{\Sigma}_\lambda$ does not impact the COS weighting scheme (2.1.2) in the sense that $\boldsymbol{w}_\lambda = \Psi(\boldsymbol{\Sigma}_\lambda) = \Psi(\widetilde{\boldsymbol{\Sigma}}_\lambda)$; similar expressions hold for COS$^+$. The equivalence between COP and COS schemes simply results from the fact that in this specific case, $\widetilde{\boldsymbol{\Sigma}}_\lambda$ takes the same form as $\boldsymbol{\Sigma}_\delta$ in (2.5) for the considered expression of δ. The equivalence holds exactly $\forall\lambda \in [0, 1)$. The equivalence also holds in the limiting case since both weighting schemes converge to equal weights, i.e.,

$$\delta \overset{\lambda\to 1}{=} +\infty \quad \text{and} \quad \boldsymbol{w}_\delta \overset{\lambda\to 1}{=} \overline{\boldsymbol{w}}^E = \boldsymbol{w}_\lambda \ . \tag{2.9}$$

Proposition 2.1.3 has an important implication for the choice of the penalty strength δ. In fact, expressing $\delta = g(\lambda)$ extends the use of an estimated shrinkage intensity in the COS with reference $\overline{\boldsymbol{\Sigma}}^E$ to the COP (with L2 divergence and with equal weights as reference scheme), and vice versa.

[7]The $\lambda = 0$ case corresponds to $\delta = 0$, i.e., no shrinkage.

2.1.6 Estimation of penalty strength

The estimate of $\boldsymbol{w}_\delta^*$ in Def. 2.1.1 depends on the chosen divergence measure $d(\cdot,\cdot)$, the reference scheme $\overline{\boldsymbol{w}}$ and the penalty strength δ. One option for selecting δ is to rely on standard k-fold cross-validation, but more advanced methods such as the forward-looking cross-validation technique used in Diebold and Shin (2019), could be considered as well. Cross-validation can mitigate the consequences of a subjective choice of $d(\cdot,\cdot)$ and $\overline{\boldsymbol{w}}$. Once δ has been found, we simply solve (2.4) by imposing the KKT conditions.

However, cross-validation forces us to choose how many observations to retain for validation, which implicitly reduces the information set that can be used for the estimation of $\boldsymbol{\Sigma}$, thereby potentially amplifying estimation error. In fact, when working with small samples, few additional observations can make the difference between singularity and more stable estimates of $\boldsymbol{\Sigma}$. In this regard, Varoquaux (2018) shows that k-fold cross-validation can strongly underestimate the variance of the prediction errors in small samples, causing forecasters to "overshrink" towards the reference weighting scheme. Hence, a closed-form expression for the penalty strength would be highly desirable.

We provide this contribution by relying on the COP-COS equivalence mentioned in Proposition 2.1.3. Following Ledoit and Wolf (2004a), we estimate the optimal shrinkage intensity λ^* by minimizing the expected value of the Frobenius norm of the difference between the shrinkage estimator of the covariance matrix and the true covariance matrix of prediction errors $\boldsymbol{S}$, i.e.,

$$\lambda^* := \underset{\lambda \in [0,1]}{\arg\min}\, \mathbb{E}\left\| (1-\lambda)\,\boldsymbol{\Sigma} + \lambda\overline{\boldsymbol{\Sigma}} - \boldsymbol{S} \right\|_2^2 . \tag{2.10}$$

with solution (Ledoit and Wolf, 2003):

$$\lambda^* = \frac{\pi - k}{\gamma} , \tag{2.11}$$

$$\begin{aligned}
\pi &= \sum_{i=1}^{m}\sum_{j=1}^{m} Var\left(\Sigma_{i,j}\right) , \\
k &= \sum_{i=1}^{m}\sum_{j=1}^{m} Cov\left(\Sigma_{i,j},\ \overline{\Sigma}_{i,j}\right) , \\
\gamma &= \sum_{i=1}^{m}\sum_{j=1}^{m} Var\left(\Sigma_{i,j}\right) + \left(\overline{\Sigma}_{i,j} - S_{i,j}\right)^2 .
\end{aligned}$$

By denoting the asymptotic variance and the asymptotic covariance as *AsyVar* and *AsyCov*, respectively [8] and using the assumptions of i.i.d. data and finite fourth moments, is a consistent estimator of λ^* (Ledoit and Wolf, 2003):

$$\widehat{\lambda}^* = \frac{1}{n}\frac{\widehat{\pi} - \widehat{k}}{\widehat{\gamma}} , \tag{2.12}$$

[8] We refer to A2.4 for the explicit formula to compute $\widehat{\lambda}^*$. The consistent estimators of $\widehat{\pi}, \widehat{k}$ and $\widehat{\gamma}$ are derived using the linear shrinkage estimator of the covariance matrix described in Ledoit and Wolf (2003, 2004a,b) and adapting the reference matrix and the notation to our framework.

$$\widehat{\pi} = \sum_{i=1}^{m}\sum_{j=1}^{m} AsyVar\left(\sqrt{n}\Sigma_{i,j}\right) ,$$
$$\widehat{k} = \sum_{i=1}^{m}\sum_{j=1}^{m} AsyCov\left(\sqrt{n}\Sigma_{i,j}, \sqrt{n}\ \overline{\Sigma}_{i,j}\right) ,$$
$$\widehat{\gamma} = \sum_{i=1}^{m}\sum_{j=1}^{m} \left(\overline{\Sigma}_{i,j} - \Sigma_{i,j}\right)^2 .$$

Clearly, the optimal shrinkage intensity λ^* depends on the reference $\overline{\mathbf{\Sigma}}$ that we have selected. Indeed, while the $\widehat{\pi}$ and $\widehat{k}$ measure respectively the instability of the elements of the sample covariance matrix[9] and the covariances between the elements of the reference and the sample covariance matrices[10], $\widehat{\gamma}$ measures the degree of misspecification of $\overline{\mathbf{\Sigma}}$. For example, if we impose the reference matrix in (2.7), $\overline{\mathbf{\Sigma}} = \overline{\mathbf{\Sigma}}^E$, then a completely inappropriate choice for the reference $\overline{\sigma}$ can easily neutralize the COS by coaxing $\lambda^* = 0$. However, by setting $\overline{\sigma}$ as in (2.7), we avoid running into this problem since they assume *a priori* that the considered predictive strategies have the same performance (i.e., the same variance of the prediction error) and that possible differences are due to chance. We note $\widehat{\delta}^* := \frac{\widehat{\lambda}^*}{1-\widehat{\lambda}^*}\overline{\sigma}^2$ the penalty strength in Proposition 2.1.3 associated with $\lambda = \widehat{\lambda}^*$. [11]

2.1.7 Negative or non-negative weights

Forecast selection and the reduction in the estimation error come at the cost of misspecification if the constraint $\boldsymbol{w} \in \mathscr{W}^+$ is violated in population. Imposing non-negative weights in portfolio selection may not look like a good idea as, perhaps, short and long positions could be the appropriate solution to the portfolio optimization problem at hand (Green and Hollifield, 1992). Nevertheless, as underlined in the Introduction, imposing a non-negativity constraint on the weights might still be beneficial because it rules out extreme solutions that essentially result from estimation errors: $\boldsymbol{w} \in \mathscr{W}^+$ implies $0 \leq w_i \leq 1$ and, when dealing with the minimum-variance objective function, coincides with an $L1$-constraint on the weights; see, e.g., DeMiguel, Garlappi, Nogales, and Uppal (2009) and Jagannathan and Ma (2003).[12]

In forecasting, negative weights would arise in very specific circumstances (Bunn, 1985; Smith and Wallis, 2009). In particular, this depends on the difference in the variance of prediction errors between models, on the correlation across candidate forecasts and on the number of models with superior performance.

For example, consider a set of 20 forecasts where a subset of them significantly underperforms the others. Table 2.1 displays the minimum difference in RMSFE necessary to attribute an optimal weight of $-0.05, -0.1, -0.2$ and -0.5 to a candidate forecast for a

[9] The sum of the asymptotic variances of the elements of $\mathbf{\Sigma}$ scaled by $\sqrt{n}$.

[10] The sum of asymptotic covariances of the entries of reference matrix with the entries of the sample covariance matrix scaled by $\sqrt{n}$. Notice that the restricting the elements of $\overline{\mathbf{\Sigma}}$ to be constant and/or independent of $\mathbf{\Sigma}$, implies that $\widehat{k} = 0$. This is the case when we shrink towards equal weights.

[11] By contrast, it does not depend on the optimization problem underlying the combination scheme nor on constraints that could be imposed on the latter.

[12] Note that some authors restrict the weights to belong to some interval $[a, b]$, where the bounds are computed from the data. For example Behr, Guettler, and Miebs (2013) show that when the investment universe is large enough, one typically gets $0 \approx a < b \leq 1$.

fixed level of correlation across prediction errors. [13] The presence of negative weights seems to be beneficial the more a candidate's prediction errors are positively correlated to the ones of the other forecasts. Nonetheless, this requires inferior (and sometimes considerably worse) performance compared to the best forecasts.

Table 2.1: Differences in performances (% RMSFE) between the k superior forecasts and $m - n$ inferior forecasts needed to achieve negative weights of different magnitude. These differences in performances are computed for distinct levels of correlation among the candidate forecast errors and for a different number k of superior models.

	Minimum negative weight -5%				
ρ	**k = 1**	**k = 2**	**k = 5**	**k = 10**	**k = 15**
0.01	>100%	>100%	>100%	>100%	>100%
0.1	>100%	>100%	>100%	>100%	>100%
0.25	>100%	>100%	>100%	>100%	>100%
0.5	>100%	>100%	>100%	**27%**	**16%**
0.75	>100%	>100%	**16%**	**8%**	**5%**
0.9	**33%**	**13%**	**5%**	**3%**	**2%**

	Minimum negative weight -10%				
ρ	**k = 1**	**k = 2**	**k = 5**	**k = 10**	**k = 15**
0.01	>100%	>100%	>100%	>100%	>100%
0.1	>100%	>100%	>100%	>100%	>100%
0.25	>100%	>100%	>100%	>100%	>100%
0.5	>100%	>100%	>100%	>100%	**34%**
0.75	>100%	>100%	>100%	**13%**	**8%**
0.9	>100%	>100%	**8%**	**4%**	**3%**

	Minimum negative weight -20%				
ρ	**k = 1**	**k = 2**	**k = 5**	**k = 10**	**k = 15**
0.01	>100%	>100%	>100%	>100%	>100%
0.1	>100%	>100%	>100%	>100%	>100%
0.25	>100%	>100%	>100%	>100%	>100%
0.5	>100%	>100%	>100%	>100%	>100%
0.75	>100%	>100%	>100%	>100%	**16%**
0.9	>100%	>100%	>100%	**7%**	**5%**

	Minimum negative weight -50%				
ρ	**k = 1**	**k = 2**	**k = 5**	**k = 10**	**k = 15**
0.01	>100%	>100%	>100%	>100%	>100%
0.1	>100%	>100%	>100%	>100%	>100%
0.25	>100%	>100%	>100%	>100%	>100%
0.5	>100%	>100%	>100%	>100%	>100%
0.75	>100%	>100%	>100%	>100%	>100%
0.9	>100%	>100%	>100%	>100%	>100%

Nevertheless, negative weights generally suggest the existence of a noise forecast, i.e., a prediction that individually does not help to predict the target, but whose inclusion in the combination of forecasts is only justified by the objective of minimizing the *in-sample* variance of prediction errors. It is well known that the inclusion of such forecasts can negatively impact the stability of the combining weights (Dickinson, 1973; Bunn, 1981). Imposing the non-negativity constraint acknowledges this fact directly in the estimation step: this differs from the use of trimming methods in a pre-screening phase. We further discuss the properties of negative and non-negative weights for COP schemes in the simulation study (Section 2.2) and the empirical applications (Section 2.3).

2.2 Monte Carlo simulations

We study the behavior of the combinations of forecasts based on constrained optimization with penalty in a controlled environment. Specifically, we analyze their sensitivity to various specifications of the variance-covariance matrix of prediction errors and to the choice of the penalty function $d(\boldsymbol{w}, \overline{\boldsymbol{w}})$. We also examine the impact on the performance of the non-negative restriction $\boldsymbol{w} \in \mathscr{W}^+$ compared to $\boldsymbol{w} \in \mathscr{W}$. Table 2.2 summarizes the combination strategies under consideration.

[13] Throughout the chapter and in this simulation exercise, we assume the individual forecasts to be unbiased. In the case of biased forecasts, the bias can be corrected by reformulating the optimal forecast combination problem in the form of a linear regression problem and adding the intercept to the model as discussed in Timmermann (2006) and indicated in footnote 2. Alternatively, the minimum-variance objective function can be adjusted so as to account for a bias-variance tradeoff.

Table 2.2: List of the considered predictive algorithms and forecast combination methods. Minimization is performed over $\mathscr{W}$ except for the models with the non-negative restriction, where it is performed on $\mathscr{W}^+$. We indicate these latter with a $^+$ superscript, e.g., CO^+.

Combination methods	Acronym	Weights
Standard methods		
Simple average forecast	SA	$\overline{\boldsymbol{w}}^E$
Loss-based weighted average forecast	IL	$\overline{\boldsymbol{w}}^{IL}$
Constrained optimization	CO$^{(+)}$	$\Psi^{(+)}(\boldsymbol{\Sigma})$
partially-egalitarian LASSO	DS_egal_lasso	$\boldsymbol{w}^{DS}$
Trimmed averages		
Trimmed simple average forecast	T-SA-p	$\mathbf{1}/\text{card}(\mathscr{M}_p)$ if $i \in \mathscr{M}_p$, 0 otherwise.
Trimmed simple average forecast via cross-validation	T-SA-CV	$\mathbf{1}/\text{card}(\mathscr{M}_p)$ if $i \in \mathscr{M}_p$, 0 otherwise.
Trimmed-MCS simple average forecast	T-MCS-α	$\mathbf{1}/\text{card}(\mathscr{M}_\alpha)$ if $i \in \mathscr{M}_\alpha$, 0 otherwise.
Constrained optimization with shrinkage direction equal weights		
COP with L1 penalty	COP1_E$^{(+)}$	$\arg\min \boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w} + \delta \left\|\boldsymbol{w} - \overline{\boldsymbol{w}}^E\right\|_1$
COP with L2 penalty	COP2_E$^{(+)}$	$\arg\min \boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w} + \delta \left\|\boldsymbol{w} - \overline{\boldsymbol{w}}^E\right\|_2^2$
COP with EN penalty	COPEN_E$^{(+)}$	$\arg\min \boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w} + \delta \left[a \left\|\boldsymbol{w} - \overline{\boldsymbol{w}}^E\right\|_1 + (1-a) \left\|\boldsymbol{w} - \overline{\boldsymbol{w}}^E\right\|_2^2\right]$
COP with EN penalty and shrinkage towards 0	COPEN	$\arg\min \boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w} + \delta \left[a \left\|\boldsymbol{w}\right\|_1 + (1-a) \left\|\boldsymbol{w}\right\|_2^2\right]$
COP with cross-entropy	COPE_E$^+$	$\arg\min \boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w} + \delta \sum_{j=1}^m w_j \ln\left(\frac{w_j}{1/m}\right)$
COS with reference $\overline{\boldsymbol{\Sigma}}^E \equiv$ COP with L2 penalty and $\widehat{\delta}^*$	COPS_E$^{(+)}$	$\Psi^{(+)}(\boldsymbol{\Sigma}_{\widehat{\delta}^*}) = \Psi^{(+)}(\boldsymbol{\Sigma}^E_{\widehat{\lambda}^*})$
Constrained optimization with shrinkage direction inverse loss-weights		
COP with L1 penalty	COP1_IL$^{(+)}$	$\arg\min \boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w} + \delta \left\|\boldsymbol{w} - \overline{\boldsymbol{w}}^{IL}\right\|_1$
COP with L2 penalty	COP2_IL$^{(+)}$	$\arg\min \boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w} + \delta \left\|\boldsymbol{w} - \overline{\boldsymbol{w}}^{IL}\right\|_2^2$
COP with EN penalty	COPEN_IL$^{(+)}$	$\arg\min \boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w} + \delta \left[a \left\|\boldsymbol{w} - \overline{\boldsymbol{w}}^{IL}\right\|_1 + (1-a) \left\|\boldsymbol{w} - \overline{\boldsymbol{w}}^{IL}\right\|_2^2\right]$
COP with cross-entropy	COPE_IL$^+$	$\arg\min \boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w} + \delta \sum_{j=1}^m w_j \ln\left(w_j/\overline{w}_j^{IL}\right)$
COS with reference $\overline{\boldsymbol{\Sigma}}^{IL}$ and $\widehat{\lambda}^*$	COS_IL$^{(+)}$	$\Psi^{(+)}(\boldsymbol{\Sigma}^{IL}_{\widehat{\lambda}^*})$
We denote with the superscript + the CO, COP1, COP2 and COPEN schemes with non-negative restriction, $\boldsymbol{w} \in \mathscr{W}^+$		

Notes. We denote the cardinality of the set $\mathscr{A}$ as $\text{card}(\mathscr{A})$. For example, $\mathscr{M}_p$ is the subset of forecasts containing $m-p$ candidates, such that $\text{card}(\mathscr{M}_p) = m-p$. We denote the trimmed simple average forecast as T-SA-p once the worst p models have been excluded. We consider the possibility of estimating the number of models to discard via cross validation in T-SA-CV. $\mathscr{M}_\alpha$ is the subset of forecasts that are included in the superior set of models with confidence level α (Hansen, Lunde, and Nason, 2011). We denote the simple average forecast derived from the candidate forecasts included in $\mathscr{M}_\alpha$ as T-MCS-α. Hence, T-MCS-α represents the simple average forecast that includes only the candidates having at least a probability α of belonging to the superior set of models. We chose α ranging from 5% to 50%. For the elastic net divergence measure, we set $a = 0.8$. Following Diebold and Shin (2019), the partially-egalitarian LASSO consists of selecting the subset of forecasts to combine in the first step, and shrinking the combining weights of the selected candidates ($\boldsymbol{w}^{DS}$) towards equal weights (using L1 divergence measure). We estimate the penalties via 5-fold cross validation both in the first and second step.

In the following sections, we consider several COP methods (Definition 2.1.1), namely, COP1, COP2, COPEN and COPE$^+$, which are associated with L1, L2, elastic net and cross-entropy divergences, respectively. The considered reference scheme $\overline{\boldsymbol{w}}$ (equal weights or inverse loss-weights) is indicated by the subscript E or IL. Minimization is performed over $\mathscr{W}$ except for the models with the non-negative restriction, where it is performed on $\mathscr{W}^+$. We indicate these latter with a $^+$ superscript[14]. For example, COP2_IL$^+$ refers to the COP scheme with L2 penalty, reference scheme $\overline{\boldsymbol{w}}^{IL}$ and $\boldsymbol{w} \in \mathscr{W}^+$, whereas COP2_IL is similar except that $\boldsymbol{w} \in \mathscr{W}$. Note that in these approaches, the penalty strength δ is chosen using cross-validation.

We also consider two COS methods (Definition 2.1.2) that rely on the analytical

[14]Recall that non-negativity is always required when using COPE.

expression for the shrinkage intensity $\widehat{\lambda}^*$, given in (2.7). For instance, COS_IL$^+$ and COS_IL use $\lambda = \widehat{\lambda}^*$ and shrink towards the reference covariance matrix $\overline{\boldsymbol{\Sigma}}^{IL}$, with and without non-negativity constraint, respectively. When considering $\overline{\boldsymbol{\Sigma}}^{E}$, Proposition 2.1.3 suggests that COS_E is equivalent to COP2_E provided that $\delta = \widehat{\delta}^*$. These equivalent COP-COS models are hence referred to as COPS_E$^+$ and COPS_E.

2.2.1 Simulation designs

We study a set of $m = 20$ candidate forecasts of a target variable $y \sim \mathcal{N}(0, \sigma_y^2)$, where a subset of k forecasts exhibits superior performance. As before, we denote by $\widehat{y}_i$ the i-th forecast and by $v_i = y - \widehat{y}_i$ its prediction error. The vectors of forecasts and prediction errors are noted $\widehat{\boldsymbol{y}}$ and $\boldsymbol{v}$, respectively. We consider two scenarios for $\boldsymbol{S}$.

In *scenario a)* we form forecasts using $\widehat{y}_i = y + v_i$ where $\boldsymbol{v} \sim \mathcal{N}(0\mathbf{1}, \boldsymbol{S})$ is independent from y. To model the fact that k models exhibit superior performance, we postulate that their RMSFEs are, on average, 50% lower than the last $m - k$ ones.[15] Hence, we set $S_{i,i} = \alpha^2$ for $i \in \{1, \dots, k\}$, $S_{i,i} = (2\alpha)^2$ for $i \in \{k+1, \dots, m\}$ and $S_{i,j\neq i} = r\sqrt{S_{i,i}S_{j,j}}$. Notice that, with a 50% difference in performance, the best forecasting strategy can feature negative weights, especially when the correlation across predictions errors is considerable. Hence, we use this scenario to study whether the non-negative restriction significantly penalizes COP schemes.

In *scenario b)*, we deal with the case where $m-k$ forecasts are "pure noise". Intuitively, an optimal combination strategy is expected to identify the k superior models while assigning negligible weights to the $m - k$ uninformative forecasts. In this setup, we model the i-th forecast as $\widehat{y}_i = y + \epsilon_i$ for $i \in \{1, \dots, k\}$ and $\widehat{y}_i = \epsilon_i$ for $i \in \{k+1, \dots, m\}$ where $\mathbb{E}[\epsilon_i] = 0$, $\mathbb{E}[\epsilon_i^2] = \alpha^2$ and $\mathbb{E}[\epsilon_i \epsilon_{j\neq i}] = r\alpha^2$. In this case, the population covariance matrix of prediction errors takes the form

$$S_{i,i} = \alpha^2 + \sigma_y^2 \mathbb{1}_{\{i>k\}} \ , \quad S_{i,j\neq i} = r\alpha^2 + \sigma_y^2 \mathbb{1}_{\{\min(i,j)>k\}}$$

where $\mathbb{1}_{\{.\}}$ is the indicator function. Notice that σ_y^2 and α control the signal to noise ratio (SNR) in this simulation exercise. We set $\sigma_y^2 = \alpha = 1$, which is representative of a scenario where the SNR = 1.

In both scenario *a)* and *b)*, we consider four *designs*, based on the number of superior forecasts k. In *design 1*, only the first two candidate forecasts have superior predictive performances ($k = 2$) for *scenario a)* or must be identified for *scenario b)*; *design 2*, *design 3* and *design 4* deal with $k = 5$, $k = 10$ and $k = 15$, respectively.

We compare the considered combination schemes (Table 2.2) with the optimal forecasting strategy $\boldsymbol{w}^*$. In particular, we compute the relative distance between the root mean square forecast error (RMSFE) of the strategy under consideration and that of the benchmark $\boldsymbol{w}^*$.[16] Specifically, for the i^{th} combination scheme, we compute the performance criterion

$$\mu_i = \frac{\text{RMSFE}_i}{\text{RMSFE}_{\boldsymbol{w}^*}} - 1 \ . \tag{2.13}$$

[15] We also considered scenarios in which the k superior models provide RMSFEs that are on average either equal or 10%, 30% and 70% lower than the inferior candidates. The results are comparable and available upon request.

[16] Note that while $\boldsymbol{w}^*$ is implied by $\boldsymbol{S}$, the considered forecast combination methods only rely on the sample counterpart of $\boldsymbol{S}$, i.e., $\boldsymbol{\Sigma} := \frac{1}{n}\boldsymbol{V}^T\boldsymbol{V}$.

We simulate $n = 1030$ observations (30 for training and 1000 for testing) for 500 replications. The difference in size between the training and testing sets should not be surprising. While having 1000 observations in the testing sample provides a rather accurate measurement of predictive performance, having 30 observations in the training sample is representative of a realistic scenario where, relative to the number of candidate forecasts, few observations are retained for estimating combining weights, and most of the data are used to train the individual candidate forecasts.[17]

2.2.2 Simulation results

We now discuss how combination methods perform across the considered scenarios. Since results are comparable across designs, Figure 2.1 only includes the results of *design 1* and *design 4* while we refer to A2.5 for *design 2* and *design 3*. For each design, the left-hand side (LHS) panel refers to *scenario a)* (superior models yield RMSFEs that are 50% lower than the inferior candidates) and the right-hand side panel (RHS) to *scenario b)* (the best forecasting strategy is sparse). We refer to A2.6 for an additional proof of the consistency of COS methods in discarding the pure noise forecasts.

We sort the combination schemes (y-axis) based on their average performances across the four correlation levels, which are indicated using colors (yellow $r = 0$, orange $r = .25$, red $r = .5$ and black $r = .9$). The closer the combination strategy is to the origin of the axes, the closer its average performance to the optimal forecasting strategy $\boldsymbol{w}^*$.

Figure 2.1 points out that the performances depend on the correlation level r of prediction errors and on whether the best forecasting strategy is sparse or not. Indeed, as anticipated in Section 2.1.6, negative weights are likely to arise when $r > .90$, which penalizes the strategies that restrict combining weights to be non-negative.

In *scenario a)* (LHS panel), we confirm this fact: non-negative COP methods outperform their negative counterparts when $r \leq .50$, but underperform the latter when $r = .90$. Such an impact can be so considerable to flip over the overall ranking. For example, in *design 1*, COPS_E$^+$ and COS_IL$^+$ perform the best when $r \leq 0.5$, but rank only fourth and sixth overall due to the poor performance in high correlation case. This can also hold for unpenalized methods: in *design 3*, while CO (i.e., the combination that only minimizes the in sample variance of prediction errors) generally performs poorly, the benefits of dropping the non-negative constraint when $r = .90$ is so considerable to counteract the superior performance of CO$^+$ (same as CO but with non-negativity constraint) in the other cases. We can generalize this finding to COP methods but not to COS: the non-negativity constraint does not penalize methods that rely on a closed form expression for the shrinkage intensity. Both in *design 1* and in *design 4*, restricting the weights to be non-negative offers an advantage: COPS_E$^+$ and COS_IL$^+$ outplay their counterparts COPS_E and COS_IL.

The non-negative restriction on the combining weights offers a significant advantage when $\boldsymbol{w}^*$ is sparse. In *scenario b)*, (RHS panel) COP methods with non-negative weights always outperform their negative counterparts, both when the penalty strength is estimated via cross-validation and via the analytical formula. In this case, the correlation level is no

[17] T-MCS methods that use bootstrapping to estimate the MCS rely on 1000 replications of the training set. Methods that use a validation fold to estimate the penalty strength retain 5 observations from the training set and rely on the remaining 25 observations for estimating the covariance matrix of prediction errors.

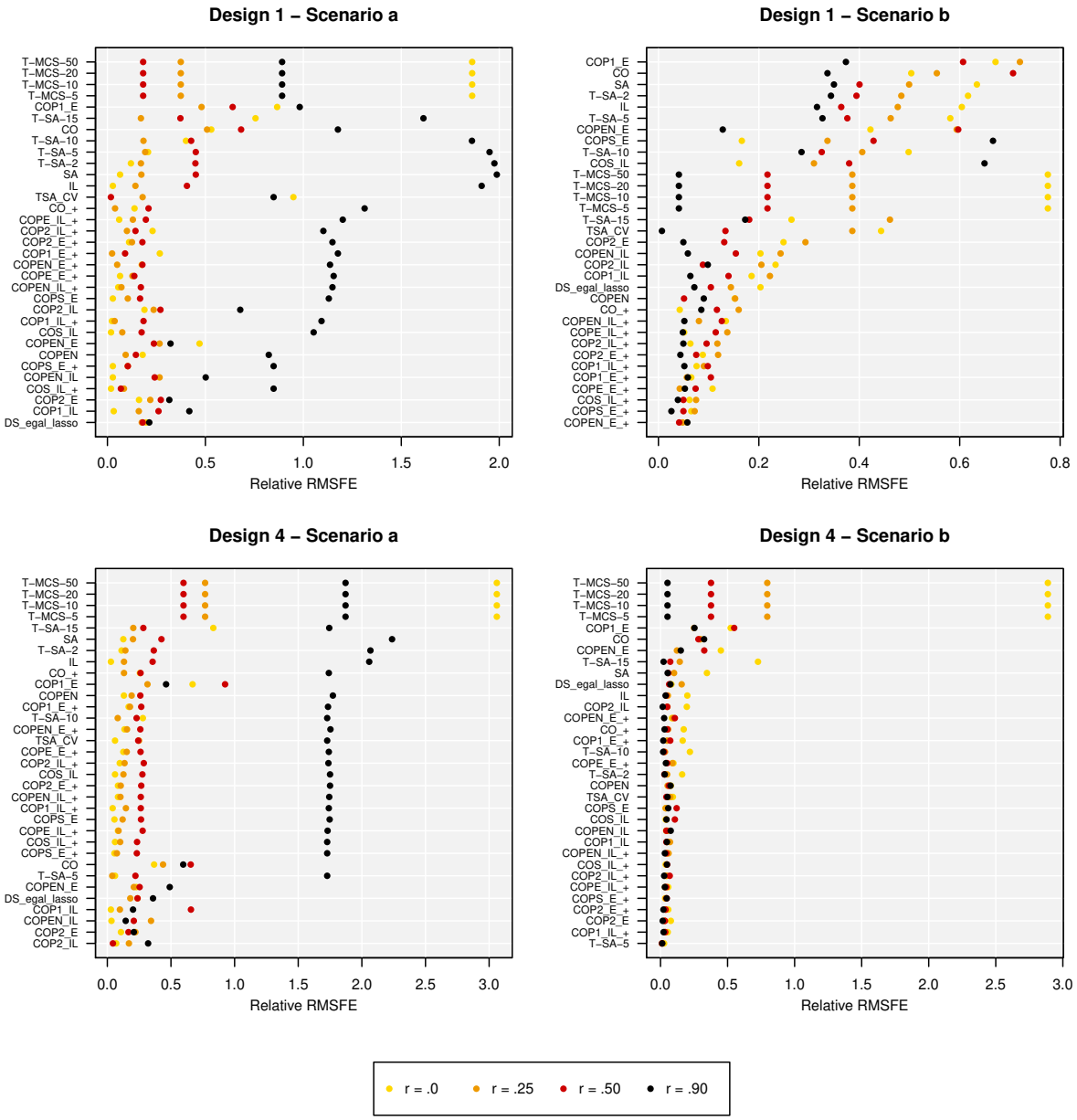


Figure 2.1: Relative RMSFE with respect to the true optimal forecast combination $\boldsymbol{w}^*$. In the left-hand side panel, superior models yield RMSFEs that are 50% lower than the inferior candidates (*scenario a*) and in the right-hand side panel the best forecasting strategy is sparse (*scenario b*). Combination schemes are ordered according to their average performances across correlation levels. In *design 1*, the first two candidates have superior predictive performance; in *design 4* the first fifteen candidates have superior performance. Right panels show the outstanding performance of the non-negativity constraint in scenario b). The same holds true for scenario a), but only for moderate correlation levels.

longer the main driver of the variability of the performance across combination strategies. Only T-MCSs remain particularly sensitive to the correlation among prediction errors.

Figure 2.1 also underlines the general poor performance of trimming methods: both

T-MCSs and T-SAs rank among the worst strategies together with IL and SA. Furthermore, estimating the number of forecasts to discard using cross-validation (T-SA-CV) does not offer any significant advantage. Only T-SA-5 performs remarkably well in the *design 4* of *scenario b)* but the reason is clear: conditional on identifying the right candidates, T-SA-5 coincides with the best forecasting strategy $\boldsymbol{w}^*$ for that design.

The partially-egalitarian LASSO procedure (DS_egal_lasso) is competitive with COP methods only when $r = .90$ and $\boldsymbol{w}^*$ is *not* sparse. We also confirm the *forecast combination puzzle*: *CO* performs poorly, especially when $r \leq .50$, where it performs worse than the simple average forecast (SA), IL and T-SAs.

This simulation study sheds the light on four aspects of great practical relevance for the use of combination strategies in forecasting.

First, all COP and COS methods perform remarkably well, especially compared to trimming-based combination strategies in all considered designs. In particular, it is difficult to identify a unique trimming method that is optimal under various specifications of $\boldsymbol{S}$. In general however, forecast selection is inappropriate when correlations among the prediction errors of candidates are particularly low ($r = 0$). Conversely, COP methods limit the impact of subjective choices on the number of models to discard.

Second, the choice of the penalty function (L1, L2, EN and cross-entropy) does not remarkably affect predictive performance when $\boldsymbol{w} \in \mathscr{W}^+$. Conversely, if $\boldsymbol{w}$ is unrestricted, the L1 divergence measure may negatively affect performance depending on the shrinkage direction ($\overline{\boldsymbol{w}}_E$ vs $\overline{\boldsymbol{w}}_{IL}$) and on the DGP under consideration.

Third, the non-negative restriction penalizes performance only when forecast errors are strongly correlated across candidate forecasts. Otherwise, in the other cases, it significantly improves predictive ability. We do not find any relevant difference between using equal weights or inverse loss-weights as reference scheme unless L1 is involved.

Fourth, COPS_E$^+$ and COS_IL$^+$ yield the minimum model risk relative to the considered combination strategies. Compared to COP1, COP2, COPEN and COPE$^+$, the analytical expressions of the shrinkage intensity in COPS_E and COS_IL allows us to make full use of the training set to estimate the covariance of prediction errors. This is a crucial advantage for more precisely estimating the combining weights.

2.3 Empirical study

In the empirical study, we consider two prediction problems taken from economics. In these applications, we train the individual algorithms using 70% of the data, build forecasts combinations on 10% and keep the remaining 20% for testing.[18] In the following sections, we present the evaluation criteria used to compare the considered predictive strategies and the predictive algorithms that used to form the candidate forecasts.

[18]This choice implies that only 21 observations are retained for constructing forecast combinations when predicting the Chicago Fed National Activity Index, and 51 observations are retained for predicting Boston housing prices. The results are robust to large estimation window, and these additional analyses are available upon request.

2.3.1 Evaluation criteria

We must define some evaluation criteria for comparing the predictive strategies. For this purpose, we consider the RMSFE[19] and two statistical procedures: the Diebold-Mariano test (hereafter DM, Diebold and Mariano, 1995) and the model confidence set (hereafter MCS, Hansen, Lunde, and Nason, 2011).

The DM tests whether one model is superior to another for a given confidence level. Formally, let the loss difference between forecast i and j for observation t be $d_{i,j}^t = (v_i^t)^2 - \left(v_j^t\right)^2$. The DM tests whether $H_0 : DM = \dfrac{\mathbb{E}[d_{i,j}]}{\sigma[d_{i,j}]} = 0$, with $DM \sim \mathcal{N}(0,1)$. However, the original test can be oversized in small samples and not robust to the heavy-tailed distributions of the forecast errors. To tackle these problems, we consider a scaled DM statistic $\sim$ standard t-student(df), as proposed by Harvey, Leybourne, and Newbold (1997).[20]

The MCS tests whether a subset of methods enters jointly in the superior set of models by repeatedly testing the null hypothesis of equal predictive performance with significance level α. Let $\mathscr{M}_0$ be the set of all forecasting models (both individual candidates and forecasts combinations), and let $\mathscr{M}^*$ be the superior set of models. Formally, the MCS tests $H_0 : \mathbb{E}\left[d_{i,j}\right] = 0,\ \forall i, j$. If the null hypothesis is rejected, then the procedure eliminates the model with the greatest relative loss from the set $\mathscr{M}$. This procedure is sequentially repeated until the null hypothesis is not rejected at the chosen probability level α. We compute the MCS p-values via bootstrapping (10000 replications) and using the Oxford MFE Toolbox publicly available at `https://www.kevinsheppard.com/code/matlab/mfe-toolbox/`.

2.3.2 Individual predictive algorithms

We consider individual forecasts from 13 standard predictive algorithms that are publicly available on software such as MATLAB. Table 2.3 lists the predictive algorithms used to form candidate forecasts.

We estimate the hyperparameters of all the considered models via 10-fold cross-validation. Ensemble methods such as gradient boosted regression tree (GBRT), random forest (RF) and feed forward neural network (NN) methods consist of 1000, 1000 and 5 base learners, respectively. This choice is an appealing compromise between computational time and out-of-sample results. Furthermore, datasets are standardized (unitary variance and zero-mean), and predictors are cross-sectionally normalized to be within $[-1, 1]$. These procedures are standard in the field of machine learning; see Gu, Kelly, and Xiu (2020) for a financial case. Combination methods that rely on validation to estimate the penalty strength use 5-fold cross-validation.

[19] We transform the target variable to have unitary variance given the information available in the training set. Therefore, RMSFE and RMSPFE (RMS-Percentage-FE) display comparable results unless the variance of the y variable significantly differs in the testing set. We present both the RMSFE and RMSPFE in A2.7.

[20] When evaluating linear forecast combinations, we may incur in the problem of testing for equal forecast accuracy for nested models. Results are robust when we also consider the correction for nested models of the DM test provided by Clark and McCracken (2001).

Table 2.3: List of the considered predictive algorithms.

Predictive model	Abbreviation	MATLAB function
Generalized linear models		
Ordinary least squares	OLS	fitlm
LASSO	L1	lasso
Ridge	L2	lasso
Elastic net	EN	lasso
Linear SVM with L1 regularization	svmL1	fitrlinear
Linear SVM with L2 regularization	svmL2	fitrlinear
Partial least squares	PLS	plsregress
Regression trees		
Gradient boosted trees	GBRT	fitrensemble
Random forests	RF	fitrensemble
Feed-forward neural networks (# of neurons per layer)		
1 Hidden layer (8)	NN1	fitnet
1 Hidden layer (16)	NN2	fitnet
2 Hidden layers (8 - 4)	NN3	fitnet
2 Hidden layers (16 - 8)	NN4	fitnet

2.3.3 Empirical study I: Boston housing prices

In the first application, Boston housing prices are predicted using 14 attributes (e.g., per capita crime rate, average number of rooms per dwelling or % lower status of the population). This dataset consists of 507 observations, and it is freely available at `http://lib.stat.cmu.edu/datasets/boston`. Although it was originally studied to provide quantitative estimates of the willingness to pay for air quality (Harrison and Rubinfeld, 1978), it is commonly used in the literature to benchmark predictive algorithms (Breiman and Friedman, 1985; Quinlan, 1993; Ben-Porat and Tennenholtz, 2017).

Figure 2.2 presents RMSFE for different combination strategies classified in two-step combinations, cross-validation-based COP methods and one step combining strategies. In this latter group, we also include the performance of the *ex post* best candidate forecasting. The vertical dotted line marks the threshold to beat before joining the superior set of models at 10% significance level. We refer the DM statistics to A2.7.

In Figure 2.2, we notice that despite using standard algorithms, our results are competitive with those of other predictive exercises on the same dataset (average RMSFE $\approx .36$[21], see `https://www.kaggle.com/c/boston-housing/leaderboard`). Nevertheless, the benefit of combining forecasts, especially via COP and COS methods, remains clear. Indeed, COPS_E, COS_IL (both with and without the non-negative weights), COP2_E$^+$, COP2_IL$^+$ and COPE_IL$^+$ rank as the best forecasting strategies. COP2_IL, COPE_IL$^+$ and the two-step partially-egalitarian LASSO of Diebold and Shin (2019) stand on the edge of the superior set of models.

Among COP and COS schemes, the L2 penalty offers the best predictive performance, while L1 performs the worst. This result does not depend on the reference combination

[21] In our application the data are standardized; therefore, to compare the our results to the leaderboard, we compute the average of the mean square forecast error of the 20 best models in the leaderboard, which gives ≈ 10.88. Then, we compute $\sqrt{10.88/84.58} = 0.3586 \approx 0.36$, where 84.58 corresponds roughly to the variance of the target variable in the data set.

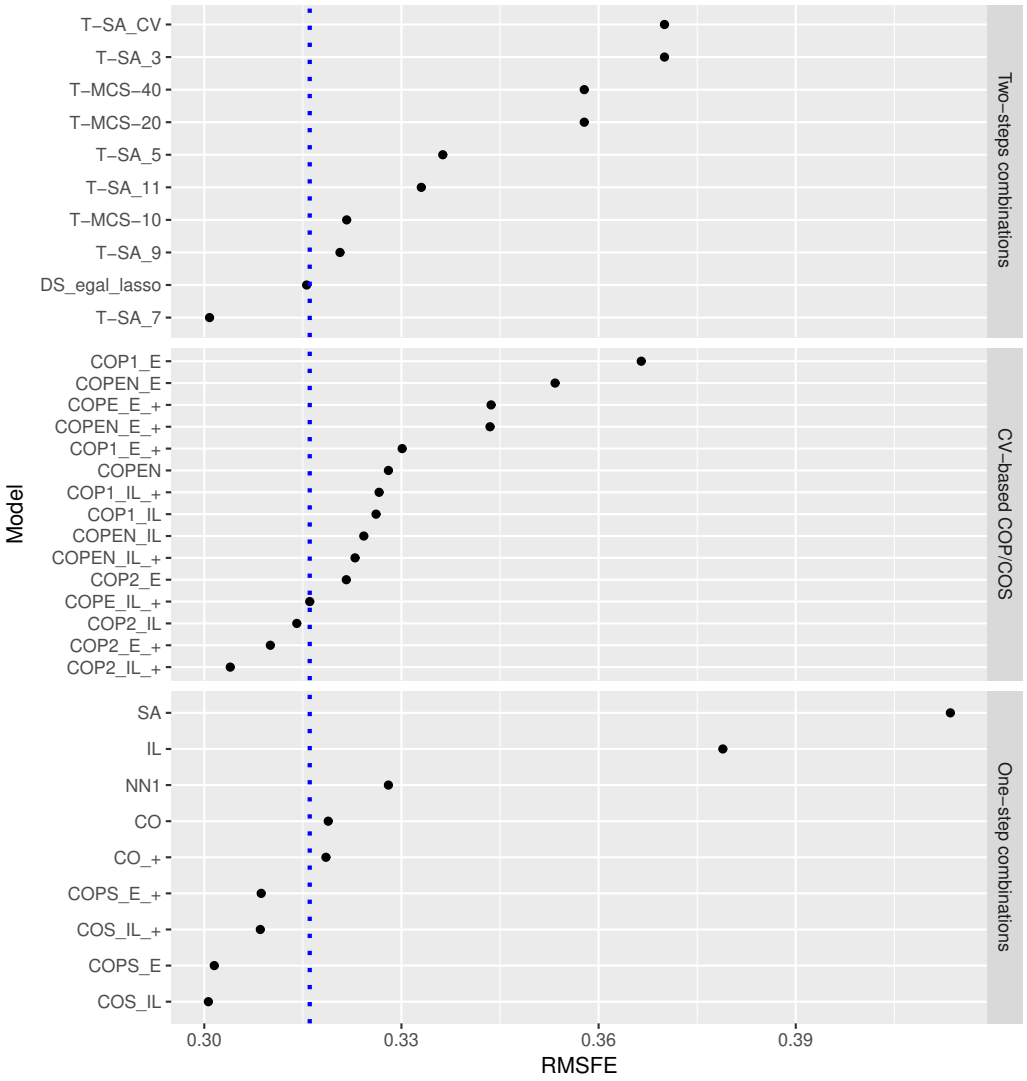


Figure 2.2: Predicting Boston housing prices, out of sample results. Performance of the forecasts combination schemes grouped in two-step combinations, cross-validation based COP methods and one step combining strategies. The vertical dotted line marks the threshold to overcome to join the superior set of models at 10% significance level.

strategy. Thus, shrinking towards equal weights or inverse loss-weights does not impact performance. However, this has one exception: COPE_E$^+$ performs significantly worse than COPE_IL$^+$.

With respect to T-SAs, the difference in predictive performance strongly depends on how many candidate forecasts we discard in the trimming phase. The findings of Aiolfi and Favero (2005) and Granger and Jeon (2004) are confirmed: among trimmed simple average strategies, those discarding most of the candidate forecasts perform best. Nevertheless, we cannot identify a clear pattern regarding the optimal number of models to discard. For example, while T-SA-7 ranks among the best performers and it is included with at least 10% in the superior set of models, T-SA-3,5,9 and 11 perform significantly worse and have less than 10% probability of being included. T-MCSs perform poorly and strongly depend on the choice of the significance level in the model confidence set. Overall, the

choice of the number of candidate forecasts to consider in T-SAs and T-MCSs is subjective. Conversely, COP methods do not require any preliminary decision about whether to trim and by how much: they only rely on the estimated covariance matrix of prediction errors to perform selection and combination in a single step.

Table 2.4: Predicting Boston housing prices, combination weights.

	OLS	L1	L2	svmL1	svmL2	EN	PLS	RF	GBRT	NN1	NN2	NN3	NN4
DS_egal_lasso	.	.	.	.	.	.	.	.25	.25	.25	.25	.	.
CO	.90	.77	(2.23)	(.23)	(.23)	.02	1.00	.40	.15	.27	.27	.15	(.23)
COPEN	.08	.09	.	.	.	.	.27	.18	.16	.09	.08	.12	(.06)
COP1_E	.08	.08	.05	(.06)	.08	.08	.08	.08	.22	.09	.08	.08	.08
COP2_E	.08	.08	.08	(.05)	(.05)	(.05)	.08	.08	.20	.20	.20	.08	.08
COPEN_E	.38	.71	(1.31)	(.54)	(.29)	.03	1.	.42	.17	.10	.15	.18	.01
COPS_E	.14	.15	(.04)	(.27)	(.26)	(.05)	.35	.25	.20	.16	.16	.19	.02
COP1_IL	.03	.03	.03	.03	.03	.03	.03	.09	.09	.15	.15	.15	.15
COP2_IL	.09	.09	(.04)	(.04)	(.04)	(.04)	.09	.09	.21	.21	.21	.09	.09
COPEN_IL	.05	.08	.02	.05	(.36)	(.30)	(.02)	.50	.26	.19	.14	.15	.13
COS_IL	.12	.13	(.03)	(.22)	(.21)	(.05)	.27	.23	.20	.17	.20	.22	(.04)
CO^+	.03	.	.	.04	.	.	.03	.18	.36	.14	.	.11	.11
$COP1\_E^+$	.07	.07	.	.	.	.07	.	.08	.20	.14	.14	.08	.14
$COP2\_E^+$	.08	.	.	.	.	.	.	.14	.21	.27	.15	.15	.
$COPE\_E^+$	.06	.06	.06	.03	.03	.03	.03	.09	.21	.15	.09	.09	.09
$COPEN\_E^+$	.04	.04	.04	.04	.04	.04	.04	.10	.15	.13	.13	.10	.10
$COPS\_E^+$	.02	.03	.	.	.	.	.02	.15	.25	.31	.	.11	.11
$COP1\_IL^+$	.05	.05	.05	.02	.02	.02	.02	.08	.14	.14	.14	.14	.14
$COP2\_IL^+$	.04	.	.	.	.	.	.	.13	.19	.19	.19	.19	.07
$COPE\_IL^+$	.03	.03	.03	.03	.03	.03	.03	.09	.12	.15	.15	.15	.15
$COPEN\_IL^+$	.02	.02	.02	.02	.02	.02	.02	.14	.14	.14	.14	.14	.14
$COS\_IL^+(r=0)$	.02	.03	.	.	.	.	.02	.16	.25	.31	.	.11	.11

Notes. We denote with $\cdot$ candidates that have been discarded and with $(\cdot)$ negative weights. Weights do sum to 1 before rounding, differences are due to the approximation to two decimal digits to enhance readability.

Table 2.4 reports the combining weights and confirms that the weights are constrained to be non-negative for forecast selection. Yet, the use of different divergence measures for COP implies different optimal combination schemes. For example, the "sharp" edges of the L1 penalty tends to force some of the weights to be exactly equal to $1/k$. Conversely, L2 better tolerates small deviations from the reference and increases weights variability, which in this case results in better predictive performance.

We also notice that, among COP methods that rely on cross-validation, only $COP1\_E^+$, $COP2\_E^+$ and $COP2\_IL^+$ provide sparse solutions: the non-negative restriction plays a crucial role in this. Also notice, that while $COPS\_E^+$ simply discards the candidates that display negative weights in COPS_E, we cannot make the same analogy for cross-validation-based methods. In the latter case, the performance in the validation fold depends on the estimated weights that clearly depend on the the non-negativity restriction. Conversely, $COPS\_E^+$, COPS_E and COS_IL^+, COS_IL have respectively identical penalty strengths, since the estimation of $\widehat{\lambda}^*$ is independent from the estimation of the weights.

However, although the weights can differ between COP methods, they largely agree on the relevance of the candidate forecasts: linear models are either discarded or play a marginal role in the aggregate forecast. In contrast, NNs, GBRT and RF are always included, and are the main drivers of the aggregate forecast. We also confirm the particular

behavior of the DS_egal_lasso: it selects a small subset of the candidates and it sets their weights to equality.

2.3.4 Empirical study II: Chicago Fed National Activity Index

The second application aims to predict the Chicago Fed National Activity Index (CFNAI), a monthly index measuring the current economic activity in the United States. The index, publicly available at `https://fred.stlouisfed.org/series/CFNAI`, is a weighted average of 85 monthly indicators of national economic activity (e.g., production and income, employment, personal consumption, housing, sales and inventories). The CFNAI is constructed to have an average value of zero and a standard deviation of one. Zero values indicate that the US economy is expanding at its historical trend rate of growth, negative values indicate below-average growth, and positive values indicate above-average growth.

We predict 1-step ahead values of CFNAI using a large dataset of backward-looking and forward-looking variables. We retrieve the backward-looking variables from the monthly dataset of the Federal Reserve Bank of St. Louis (FRED-MD) and transform the original series according to the work of McCracken and Ng (2016). We also include measures of macroeconomic, financial and real uncertainty from Jurado, Ludvigson, and Ng (2015b) and Ludvigson, Ma, and Ng (2015). We retrieve the forward-looking variables from the black Chip Financial Forecasts (BCFF) database. This database relies on a survey of professional forecasters that includes the expected term structure of U.S. interest rates, expected real GDP, price level (GDP price deflator and Consumer Price Index) and the expected level of the major currency index. BCFF also includes the average expected yield for home mortgages, and the yield of Aaa- and Bbb-rated corporate bonds. We use the first three principal components to summarize the information of the term structure of expected yields. We further compute the standard deviation and the conditional asymmetry (Bowley, 1920) to capture the dispersion and the skewness of the empirical distribution of forecasts, respectively. We count 125 predictive variables. Our sample starts in January 2001 and stops in May 2018 (209 monthly observations). We train the individual models using 146 observations, form a combination of forecasts using 21 observations and keep the remaining 42 observations for testing. Thus, COPS_E and COS_IL have $21 - 13 = 8$ degrees of freedom, while methods that rely on 5-fold cross-validation have only $17 - 13 = 4$ degrees of freedom for computing $\mathbf{\Sigma}$.

Note that in this predictive exercise, the time dimension can affect the predictive ability of the individual models and, consequently, combination methods. Therefore, as a robustness check, we also consider an expanding window to estimate the individual models, to form the combination of forecasts and to form evaluation statistics. This means that we estimate the individual models and the combining weights 42 times. The training set, made of the data available up to time t, is partitioned in two folds. The *combination fold* is composed of 21 observations selected randomly from the training set, and is used to estimate the combining weights. The other fold is used to train the individual models. Results are robust to this second specification and we refer to A2.8 for the results that build on the expanding window specification.

Figure 2.3 presents the obtained RMSFEs; the Diebold-Mariano statistics are provided in the Appendix. The results are in line with the previous findings.

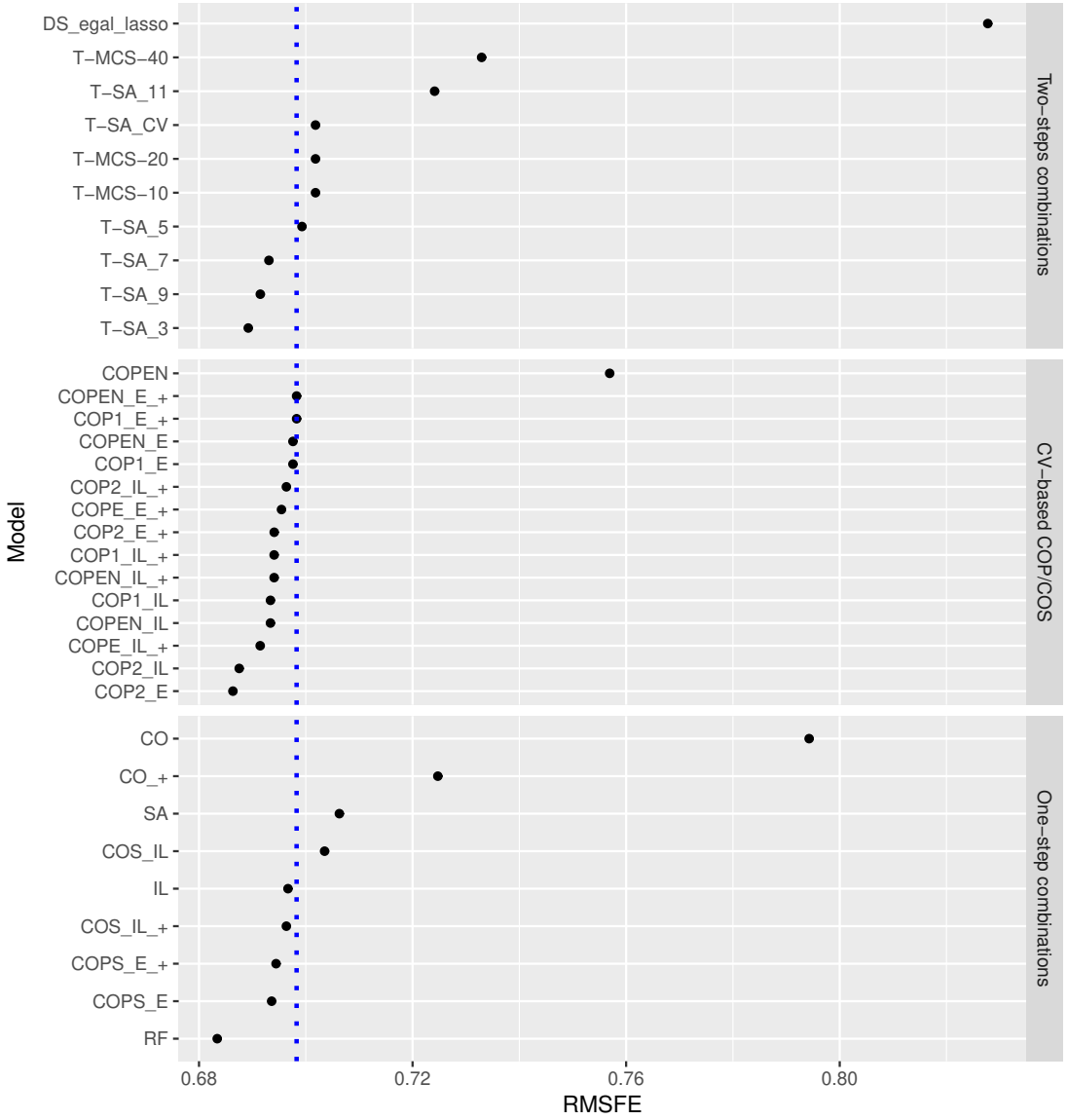


Figure 2.3: Predicting Chicago Fed National Activity Index (CFNAI), out of sample results. Performance of the forecasts combination schemes grouped across two-step combinations, cross-validation based COP methods and one step combining strategies. The vertical dotted line marks the threshold to overcome to join the superior set of models at 10% significance level.

In Figure 2.3, we remark that the COP and COS methods significantly outperform our main benchmarks, i.e., SA, CO^+ (Conflitti, Mol, and Giannone, 2015) and DS_egal_lasso (Diebold and Shin, 2019). The use of a single step method with penalization is what marks the difference between being excluded (SA, CO, CO^+ and DS_egal_lasso) and being included in the superior set of models (COP and COS schemes). On the one hand, despite imposing the non-negative constraint on the weights improves the performance (CO^+ outperforms CO), this restriction alone is not sufficient to be included in the superior set of models (CO^+ is excluded). On the other hand, the partially-egalitarian LASSO

(DS_egal_lasso) actually collapses to NN3 (see Table 2.5), which however is not the best candidate forecast in the testing set: this underlines how the failure to identify the model(s) in the first step can have relevant consequences on performance. We can draw similar conclusions on T-SAs and T-MCSs.

T-SAs are competitive with the best forecasting strategies, but it remains difficult to clearly determine the *right* number of models to discard: while TSA_3,_9 and TSA_7 are included in the superior set of models, TSA_5 and TSA_11 are excluded. As stressed earlier, a wrong choice for this number can be very harmful: TSA_11 ranks among the worst performing strategies overall. T-MCS methods are always excluded from the superior set of model. Conversely, COS$^+$ schemes are always included in the superior set of models and, at 10% significance level, their performance are not statistically different from the best *ex post* forecasting strategy, i.e., RF.

Table 2.5 gives an overview of the estimated weights. We find evidence that COP methods can achieve sparsity; i.e., they combine only a subset of the available candidate forecasts. Furthermore, as in the previous application, COP and COS methods and largely agree on which models to select and on the weight to assign to each of them. Interestingly, we notice that COS_IL, COP2_IL and COPS_E, COP2_E respectively estimate almost identical combining weights. Furthermore, the non-negativity restriction does not play any role in COP1_E$^+$, COP1_IL$^+$ and COPEN_IL$^+$ because the corresponding weights are positive even when the non-negativity constraint is relaxed.

Table 2.5: Predicting Chicago Fed National Activity Index (CFNAI), combination weights.

	OLS	L1	L2	svmL1	svmL2	EN	PLS	RF	GBRT	NN1	NN2	NN3	NN4
DS_egal_lasso	.	.	.	.	.	.	.	.	.	.	.	1.	.
CO	(.09)	(.93)	.01	.99	1.	(.46)	.55	.10	(.18)	(.67)	(.02)	1.	(.31)
COPEN	.	.	.	1.	.	.	.	.	.	.	.	.	.
COP1_E	.05	.08	.08	.11	.08	.08	.08	.08	.08	.08	.08	.08	.08
COP2_E	(.03)	.03	.04	.14	.12	.04	.04	.16	.06	.06	.05	.23	.06
COPEN_E	.05	.08	.08	.11	.08	.08	.08	.08	.08	.08	.08	.08	.08
COPS_E	(.08)	(.09)	(.08)	.34	.48	(.20)	.02	.49	(.17)	(.17)	(.17)	.90	(.26)
COP1_IL	.01	.08	.08	.10	.10	.08	.08	.08	.08	.08	.08	.08	.07
COP2_IL	(.04)	.03	.04	.15	.13	.04	.04	.16	.05	.06	.06	.23	.05
COPEN_IL	.01	.08	.08	.10	.10	.08	.08	.08	.08	.08	.08	.08	.07
COS_IL	(.08)	(.10)	(.11)	.38	.60	(.24)	.01	.49	(.19)	(.20)	(.20)	.92	(.28)
CO$^+$	.	.	.	.54	.33	.	.	.	.	.13	.	.	.
COP1_E $^+$	.05	.08	.08	.11	.08	.08	.08	.08	.08	.08	.08	.08	.08
COP1_IL$^+$	.01	.08	.08	.10	.10	.08	.08	.08	.08	.08	.08	.08	.07
COP2_E$^+$	.	.	.	.27	.26	.	.	.24	.	.10	.	.13	.
COP2_IL$^+$	.	.	.	.31	.27	.	.	.21	.	.09	.	.13	.
COPE_E$^+$	.05	.08	.08	.11	.08	.08	.08	.08	.08	.08	.08	.08	.08
COPE_IL$^+$	.01	.08	.08	.10	.10	.08	.08	.08	.08	.08	.08	.08	.07
COPEN_E$^+$	.03	.07	.07	.09	.09	.08	.07	.10	.08	.07	.07	.10	.07
COPEN_IL$^+$	.	.07	.08	.11	.10	.08	.07	.10	.07	.08	.08	.11	.07
COPS_E$^+$	.	.	.	.26	.24	.	.	.25	.	.08	.05	.13	.
COS_IL$^+(r=0)$	.	.	.	.30	.27	.	.	.21	.	.09	.	.13	.

Notes. We denote with $\cdot$ candidates that have been discarded and with $(\cdot)$ negative weights. Weights do sum to 1 before rounding, differences are due to the approximation to two decimal digits to enhance readability.

2.4 Conclusions

Constrained optimization with penalty and shrinkage techniques have been successfully used in finance, and they are at the heart of asset allocation strategies. In this chapter, we point out that these methods also provide an appealing route for forming forecast combinations. In particular, we design a robust single-step framework that allows to obtain *potentially sparse* combining weights. We show that these methods, relying on constrained optimization with penalty (COP) or with shrinkage of the covariance matrix (COS) and the non-negativity constraint on the combining weights, limit the subjectivity in the trimming phase because they do not require any predetermined selection rule or predetermined combination strategy; they rely on the estimated covariance matrix of prediction errors.

We also present and discuss the scenarios where dropping the non-negativity restriction can be beneficial. We find that this occurs only under specific circumstances: when a candidate's prediction errors are strongly positively correlated to those of the other forecasts and when, at the same time, its predictive ability is also significantly inferior. Moreover, we introduce new divergence measures in COP (namely, elastic net and cross-entropy) and study their performance. We find promising results for cross-entropy (only valid when working under the non-negativity constraint), which consistently outperforms the standard L1 penalization. Eventually, we use the correspondence between the shrinkage estimators of the sample covariance matrix (Ledoit and Wolf, 2004a) and the L2-norm penalty (DeMiguel, Garlappi, Nogales, and Uppal, 2009) to obtain an analytical expression for the optimal penalty strength, thereby making the economy of a potentially harmful cross-validation step in this case.

In an extensive simulation study and in two empirical applications, we document the flexibility and robustness of the single-step combinations with non-negative weights and explicit penalty strength (or shrinkage intensity). We find that these models outperform cross-validation based COP methods, the simple average forecast, trimming methods and other standard benchmarks from previous studies.

APPENDIX

A2.1 Proof of Proposition 2.1.2

Specializing the optimization problem to the Euclidean distance between $\mathbf{w}$ and the equally weighted scheme yields (2.1.1),

$$\boldsymbol{w}_\delta^* := \underset{\boldsymbol{w} \in \mathscr{W}}{\arg\min}\ \boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w} + \delta \left\|\mathbf{w} - \overline{\boldsymbol{w}}^E\right\|_2^2 .$$

The penalty term is developed by making use of the fact that $\boldsymbol{w} \in \mathscr{W}^+ \subset \mathscr{W}$:

$$\|\mathbf{w} - \mathbf{1}/m\|_2^2 = \mathbf{w}^T\mathbf{w} - 1/m .$$

By discarding the constant and grouping the quadratic terms,

$$\boldsymbol{w}_\delta^* := \underset{\boldsymbol{w} \in \mathscr{W}^+}{\arg\min}\ \boldsymbol{w}^T\left(\boldsymbol{\Sigma} + \delta\mathbf{I}\right)\boldsymbol{w} ,$$

which takes a similar form as ((2.2)). The proof is concluded by applying Proposition 2.1.1, replacing $\boldsymbol{\Sigma}$ by $\boldsymbol{\Sigma} + \delta\mathbf{I}$ (observe that because $\boldsymbol{\Sigma} + \delta\mathbf{I}$ is positive definite, which is always assumed to be true, $\boldsymbol{\Sigma}$ and $\delta \geq 0$ are positive definite). Notice that this result holds for a weighting scheme $\boldsymbol{w} \in \mathscr{W}$.

A2.2 Proof of Equivalence: A general case

The general form of the reference covariance matrix associated with the equally weighted scheme can be written as $\overline{\boldsymbol{\Sigma}} = \overline{\sigma}^2\Big((1-\rho)\mathbf{I} + \rho\mathbf{E}\Big)$, where $\mathbf{E}$ is a conformable matrix of ones. It is easy to see that the optimization problem ((2.6)) associated with the specific reference matrix above collapses to that associated with $\rho = 0$ under the $\mathbf{w} \in \mathscr{W}^+$ constraint. Indeed, for every $\lambda \in [0, 1]$, the objective function can be decomposed as

$$\boldsymbol{w}^T\boldsymbol{\Sigma}_\lambda\boldsymbol{w} = \boldsymbol{w}^T\left(\boldsymbol{\Sigma} + \frac{\lambda}{1-\lambda}(1-\rho)\overline{\sigma}^2\mathbf{I}\right)\boldsymbol{w} + \frac{\lambda}{1-\lambda}\rho\overline{\sigma}^2\boldsymbol{w}^T\mathbf{E}\boldsymbol{w} .$$

Due to the restriction $\mathbf{w} \in \mathscr{W}^+$, the last term collapses to the constant $\lambda\rho\overline{\sigma}^2/(1-\lambda)$, which can be discarded. Hence, the cost function is that of a COS scheme with $\overline{\boldsymbol{\Sigma}} = \overline{\sigma}^2\mathbf{I}$ and shrinkage intensity $\frac{\lambda}{1-\lambda}(1-\rho)$. Thus, we can set $\rho = 0$ without loss of generality.

A2.3 Bayesian reformulation of the COP problem

We motivated the use of COP and COS schemes with non-negativity constraint to improve the stability of the combining weights when the sample covariance matrix is poorly estimated or when the number of candidate forecasts is greater than the number of available observations. Nevertheless, we can also motivate the use of shrinkage from a Bayesian perspective. On the one hand, it is a way to incorporate prior information into the estimation of combining weights as suggested by Diebold and Pauly (1990). On the other hand, a Bayesian reformulation of COP helps the forecaster to quantify the uncertainty

around the estimated weights and incorporate the doubts about the non-negative restriction on the combining weights.

In this section, we reformulate COP problem with L1 and L2 divergence measure in a Bayesian framework. We restrict ourselves to consider the reference scheme $\overline{\boldsymbol{w}}^E$ since Bayesian prior should not depend on the the specific sample under analysis. Following Granger and Ramanathan (1984), we recast the optimal forecast combination as a linear regression problem. Denoting by $\boldsymbol{Y} := [\boldsymbol{y}, ..., \boldsymbol{y}]$, the $n \times m$ matrix containing m copies of $\boldsymbol{y}$, and by $\boldsymbol{V}$ the $n \times m$ matrix containing the realized forecast errors, we can reformulate the CO regression problem as

$$\boldsymbol{w}^* := \underset{\boldsymbol{w}\in\mathscr{W}}{\arg\min} \frac{1}{n} \left[\boldsymbol{y} - \widehat{\boldsymbol{Y}}\boldsymbol{w}\right]^T \left[\boldsymbol{y} - \widehat{\boldsymbol{Y}}\boldsymbol{w}\right] = \underset{\boldsymbol{w}\in\mathscr{W}}{\arg\min} \frac{1}{n} \left[\boldsymbol{y} - (\boldsymbol{Y} + \boldsymbol{V})\,\boldsymbol{w}\right]^T \left[\boldsymbol{y} - (\boldsymbol{Y} + \boldsymbol{V})\,\boldsymbol{w}\right] .$$

Recalling that $\mathbf{1}^T\boldsymbol{w} = 1$, we notice that $\boldsymbol{Y}\boldsymbol{w} = \boldsymbol{y}$. Therefore,

$$\boldsymbol{w}^* := \underset{\boldsymbol{w}\in\mathscr{W}}{\arg\min} \frac{1}{n} \left[\boldsymbol{V}\boldsymbol{w}\right]^T \left[\boldsymbol{V}\boldsymbol{w}\right]^T = \underset{\boldsymbol{w}\in\mathscr{W}}{\arg\min}\, \boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w} .$$

Finally, we can retrieve COP2 by adding the appropriate penalty term. In Proposition 2.1.2, we show that when shrinkage towards equal weights, the solution to COP2_E consists of shrinking diagonal elements of $\boldsymbol{\Sigma}$ by δ. In this case, the posterior mode coincides with the Ridge estimate of $\boldsymbol{w}$ when reformulating COP2 in the form of a linear regression problem (see for example Hastie, Tibshirani, and Friedman, 2008). In particular, by using the Gaussian likelihood, the Normal-inverse-Wishart distribution as prior and appropriately rescaling the weights to sum to 1, we obtain the Bayesian reformulation of COP2_E (see Geweke, 2005, for an introduction to Bayesian regression models). Similarly, we can retrieve COP1_E by imposing a Laplace prior (centered on the reference weighting scheme) on the weights. COP1_E can be easily implemented by modifying the Gibbs sampler proposed by Park and Casella (2008) to center the Laplace prior around $\overline{\boldsymbol{w}}^E$.

Notice that we neither directly impose a truncated Gaussian (COP2_E) or a truncated Laplace (COP1_E) prior on the weights nor truncate the posterior distribution in $[0, 1]$. Indeed, despite we can quantify the uncertainty on the estimated weights also when restricting weights to be non-negative, dropping this assumption permits us to also incorporate the doubts about the non-negative restriction itself. While the underlying hypothesis is that negative weights signal noise forecasts that should be discarded as discussed in Section 2.1.1, if the posterior indicates that a weight is significantly negative, the Bayesian framework would have helped us identifying a case where this restriction is violated in population. The mode, but most importantly the dispersion of the posterior distribution of the weights, should guide the forecaster on whether including or discarding a subset of forecasts. However, dropping the non-negative restriction comes at one cost: COP1_E and COP2_E cannot perform forecast selection and therefore handle high dimensional combination problems.

A2.4 Analytical formula for the optimal shrinkage intensity

Following Ledoit and Wolf (2003, 2004b,a), we present the analytical formula to estimate the optimal shrinkage intensity for COS methods. This formula for λ^* makes the linear shrinkage estimator of $\boldsymbol{\Sigma}$ asymptotically optimal, i.e., it asymptotically minimizes the

Frobenious Norm (quadratic loss function) of the difference between the true and the shrinkage estimator $\widehat{\mathbf{\Sigma}}_\lambda$ as the number of observations and the number of variables go to infinity together. As indicated in ((2.12)), the optimal shrinkage intensity is

$$\widehat{\lambda}^* = \frac{1}{n}\frac{\widehat{\pi} - \widehat{k}}{\widehat{\gamma}} ,$$

$$\begin{aligned}
\widehat{\pi} &= \sum_{i=1}^{m}\sum_{j=1}^{m} AsyVar\left(\sqrt{n}\Sigma_{i,j}\right) , \\
\widehat{k} &= \sum_{i=1}^{m}\sum_{j=1}^{m} AsyCov\left(\sqrt{n}\Sigma_{i,j}, \sqrt{n}\ \overline{\Sigma}_{i,j}\right) , \\
\widehat{\gamma} &= \sum_{i=1}^{m}\sum_{j=1}^{m} \left(\overline{\Sigma}_{i,j} - \Sigma_{i,j}\right)^2 .
\end{aligned}$$

While $\widehat{\gamma}$ can be easily computed from ((2.12)), the formulas for $\widehat{\pi}$ and $\widehat{k}$ are less straightforward. Following the work of Ledoit and Wolf, we provide the corresponding formulas below.

Denoting the z^{th} realized forecast error of the i^{th} model and i^{th} model's sample average forecast error with $V_{z,i}$ and $V_{\cdot,i}$, respectively, we start by considering $\widehat{\pi}$, i.e., a consistent estimator of π (Lemma 1 p. 612 Ledoit and Wolf, 2003)

$$\widehat{\pi} = \sum_{i=1}^{m}\sum_{j=1}^{m} \widehat{\pi}_{ij} , \tag{A2.1}$$

$$\widehat{\pi}_{ij} = \frac{1}{n}\sum_{z=1}^{n}\left[\left(V_{zi} - \frac{1}{n}V_{\cdot i}\right)\left(V_{zj} - \frac{1}{n}V_{\cdot j}\right) - \Sigma_{ij}\right]^2 . \tag{A2.2}$$

Notice that π and its consistent estimator $\widehat{\pi}$ do not depend on the reference matrix $\overline{\mathbf{\Sigma}}$. Conversely, consider $\widehat{k}$, i.e a consistent estimator of k (Lemma 2 p. 612-613 Ledoit and Wolf, 2003).

$$\widehat{k} = \sum_{i=1}^{m}\sum_{j=1}^{m} \widehat{k}_{ij} .$$

By construction, $\widehat{k}$ depends on the reference matrix $\overline{\mathbf{\Sigma}}$. When $\overline{\mathbf{\Sigma}} = \overline{\mathbf{\Sigma}}^E = \overline{\sigma}^2 I$, the covariance each element of the sample covariance matrix and the identity matrix scaled by $\overline{\sigma}^2$ is constantly equal to zero. Therefore, in the case of COS_E$^{(+)}$, the consistent estimator of the optimal shrinkage intensity $\widehat{\lambda}^*$, is given by $\widehat{\lambda}^* = \frac{\widehat{\pi}}{\widehat{\gamma}}$. This corresponds to the optimal shrinkage intensity derived in Theorem 2.1 in Ledoit and Wolf (2004b).

A2.5 Simulation study: Design 2 and Design 3

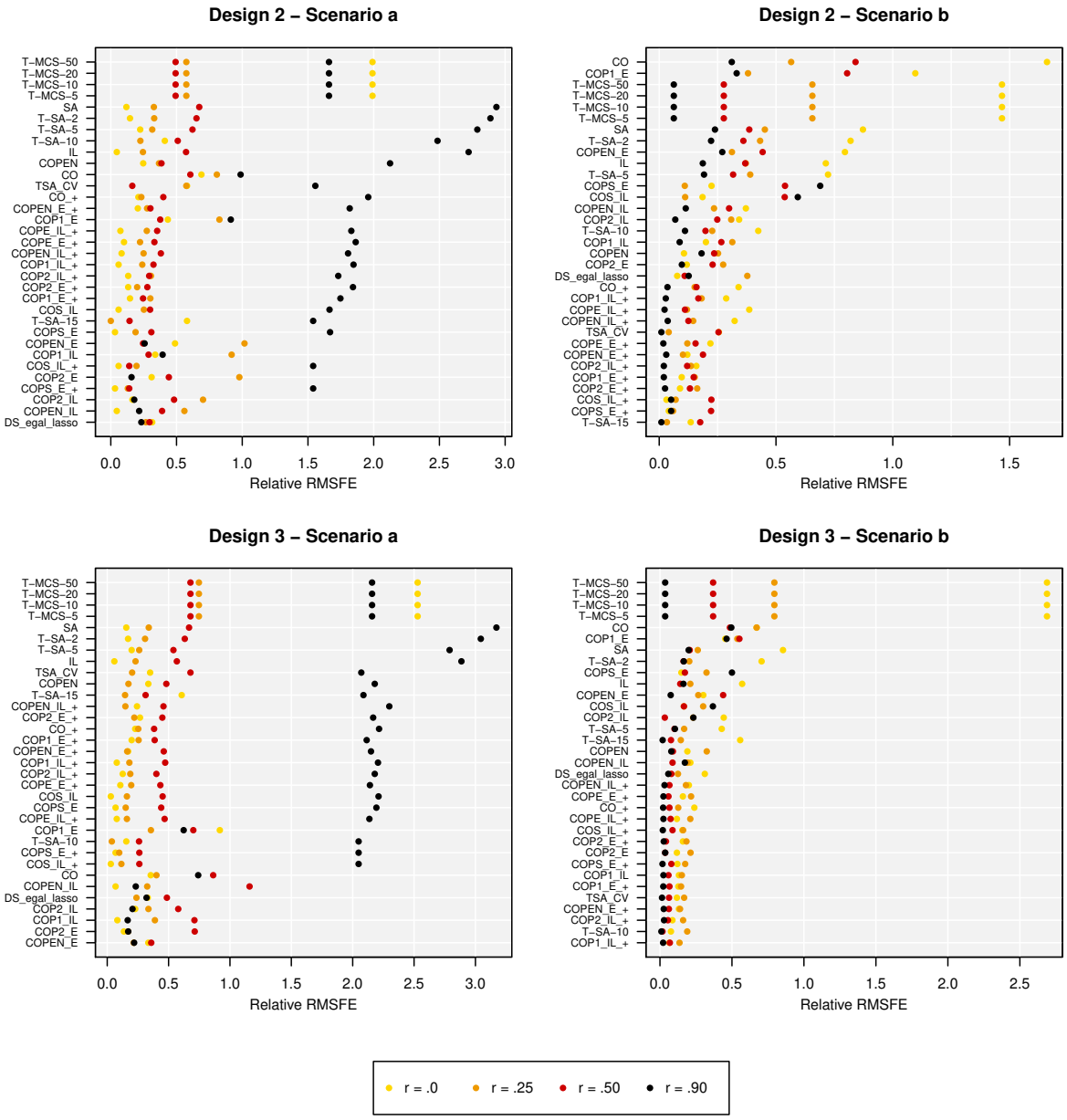


Figure A2.1: Relative RMSFE with respect to the true optimal forecast combination $\boldsymbol{w}^*$. Performance of the forecasts combination schemes for various specifications of the covariance matrix $\boldsymbol{\Sigma}_v$ of the forecasts errors. Combination schemes are ordered according to their average performances across correlation levels. In each design, the candidates that have superior predictive performance yield RMSFEs that are 50% lower than the remaining ones. In *design 2*, the first five candidates have superior predictive performance; in *design 3* the first ten candidates have superior performance.

A2.6 Forecast selection using CO+ and COS(+) methods.

Table A2.1: L2 norm of weights assigned to noise forecasts.

	r = 0	r= 0.25	r=0.5	r=0.9
T = 30		$\alpha_{noise} = 10$		
CO+	0.009	0.007	0.007	0.007
COPS_E	0.021	0.004	0.001	0.002
COS_IL	0.001	0.001	0.001	0.002
COPS_E+	0.021	0.001	0.000	0.000
COS_IL+	0.001	0.000	0.000	0.000
T = 30		$\alpha_{noise} = 1$		
CO+	0.017	0.017	0.018	0.015
COPS_E	0.020	0.031	0.046	0.083
COS_IL	0.018	0.027	0.040	0.068
COPS_E+	0.015	0.012	0.012	0.014
COS_IL+	0.013	0.011	0.011	0.013
T = 100		$\alpha_{noise} = 10$		
CO+	0.008	0.007	0.007	0.007
COPS_E	0.009	0.000	0.000	0.000
COS_IL	0.001	0.000	0.000	0.000
COPS_E+	0.009	0.000	0.000	0.000
COS_IL+	0.001	0.000	0.000	0.000
T = 100		$\alpha_{noise} = 1$		
CO+	0.015	0.015	0.015	0.011
COPS_E	0.015	0.024	0.033	0.065
COS_IL	0.014	0.023	0.032	0.061
COPS_E+	0.010	0.010	0.010	0.006
COS_IL+	0.010	0.010	0.009	0.006

Notes. L2 norm of the estimated weights assigned to noise forecasts. We model the i-th forecast as $\widehat{y}_i = y + \epsilon_i$ for $i \in \{1, \ldots, 10\}$ and $\widehat{y}_i = \alpha_{noise}\epsilon_i$ for $i \in \{11, \ldots, 20\}$ where $\mathbb{E}[\epsilon_i] = 0$, $\mathbb{E}[\epsilon_i^2] = \alpha^2 = 1$ and $\mathbb{E}[\epsilon_i \epsilon_{j \neq i}] = r\alpha_i\alpha_j^2$. Despite the reference weighting schemes $\overline{\boldsymbol{w}}^{IL}$ and $\overline{\boldsymbol{w}}^{E}$ assign non-zero weights to noise forecasts, aspect that penalizes COS' selection ability, COS methods consistently discard noise forecasts. The reason is that as the sample size increases, we also more precisely estimate $\boldsymbol{\Sigma}$: the shrinkage intensity λ^* will asymptotically decay, providing the consistency result of the linear shrinkage estimator of the covariance matrix of Ledoit and Wolf (2003). Thus, the L2 norm of the weights assigned to noise forecasts approaches zero as the sample size increases. This holds for any level of correlation.

A2.7 Performance of all models in the empirical studies

Table A2.2: Forecasting Boston housing prices, out-of-sample performance.

	Diebold-Mariano		RMSFE	RMSPFE
i \ j:	COPS-E$^+$	COS-IL$^+$		
OLS	$\lll$	$\lll$	.5639	.5713
L1	$\lll$	$\lll$	.5638	.5712
L2	$\ll$	$\ll$	.5687	.5761
svmLASSO	$\ll$	$\ll$	.5963	.6041
svmRIDGE	$\ll$	$\ll$	.5947	.6025
EN	$\lll$	$\lll$	.5952	.6030
PLS	$\ll$	$\ll$	.5845	.5922
RF	$\lll$	$\lll$	.4567	.4627
GBRT	$\lll$	$\lll$	.3763	.3812
NN1*	=	=	.3280	.3323
NN2	$\ll$	$\ll$	.3578	.3625
NN3	=	=	.3320	.3364
NN4	=	=	.3304	.3348
SA	$\ll$	$\ll$	.4135	.4183
T-SA-3	$\ll$	$\ll$	.3700	.3748
T-SA-5	$\ll$	$\ll$	.3363	.3411
T-SA-7	=	=	.3008	.3056
T-SA-9	=	=	.3206	.3254
T-SA-11	$<$	$<$	.3330	.3378
T-SA-CV	$\ll$	$\ll$	.3700	.4048
T-MCS-60%	=	=	.3217	.3264
T-MCS-80%	$\ll$	$\ll$	.3578	.3626
T-MCS-90%	$\ll$	$\ll$	.3578	.3626
IL	$\ll$	$\ll$	.3789	.3837
DS_egal_lasso	=	=	.3156	.3203
CO	=	=	.3188	.3236
COPEN	$\ll$	$\ll$	.3280	.3328
CO$^+$	=	=	.3185	.3223
COP1_E$^+$	$\ll$	$\ll$	.3301	.3349
COP2_E$^+$	=	=	.3100	.3148
COPE_E$^+$	$<$	$<$	.3437	.3484
COPEN_E$^+$	$\ll$	$\ll$	.3435	.3482
COP1_IL$^+$	$\ll$	$\ll$	.3266	.3314
COP2_IL$^+$	=	=	.3040	.3087
COPE_IL$^+$	=	=	.3160	.3208
COPEN_IL$^+$	$\ll$	$\ll$	.3230	.3277
COP1_E	$\ll$	$\ll$	.3665	.3713
COP2_E	$\ll$	$\ll$	.3216	.3263
COPEN_E	$<$	$<$	.3533	.3581
COP1_IL	$\ll$	$\ll$	.3261	.3309
COP2_IL	=	=	.3140	.3088
COPEN_IL	$<$	$<$	.3243	.3290
COPS_E	=	=	.3015	.3063
COS_IL	=	=	.3006	.3054
COPS_E$^+$	.	=	.3087	.3134
COS_IL$^+$	=	.	.3086	.3133

Notes. The symbols $<$, $\ll$ and $\lll$ indicate that model j (column index) surpasses model i (row index) at confidence levels of 90%, 95% and 99%, respectively. Failure to reject the null hypothesis of equal predictive performance at a 10% significance level is indicated with "=".

Table A2.3: Predicting Chicago Fed National Activity Index (CFNAI), out-of-sample performance.

	Diebold-Mariano		RMSFE	RMSPFE
i \ j:	COPS-E$^+$	COS-IL$^+$		
OLS	$\lll$	$\lll$	1.280	1.321
L1	=	=	.700	.723
L2	=	=	.698	.721
svmLASSO	=	=	.757	.782
svmRIDGE	=	=	.733	.757
EN	=	=	.686	.708
PLS	=	=	.718	.741
RF*	=	=	.683	.706
GBRT	=	=	.740	.764
NN1	$\lll$	$\lll$	.822	.849
NN2	$\lll$	$\lll$	.821	.848
NN3	$\ll$	$\lll$	.829	.856
NN4	$\lll$	$\lll$	.856	.884
SA	=	=	.706	.729
T-SA-3	=	=	.689	.712
T-SA-5	=	=	.699	.722
T-SA-7	=	=	.693	.716
T-SA-9	=	=	.691	.714
T-SA-11	$\ll$	$\ll$	.724	.748
T-SA-CV	=	=	.702	.725
T-MCS-60%	=	=	.702	.725
T-MCS-80%	=	=	.702	.725
T-MCS-90%	=	=	.733	.757
IL	=	=	.697	.719
DS_egal_lasso	$\ll$	$<$	.828	.855
CO	$<$	$<$	.794	.820
COPEN	=	=	.757	.782
CO$^+$	$<$	$<$	.725	.748
COP1_E $^+$	=	=	.698	.721
COP2_E$^+$	=	=	.694	.717
COPE_E$^+$	=	=	.698	.721
COPEN_E$^+$	=	=	.695	.718
COP1_IL$^+$	=	=	.694	.717
COP2_IL$^+$	=	=	.696	.719
COPE_IL$^+$	=	=	.694	.717
COPEN_IL$^+$	=	=	.691	.714
COP1_E	=	=	.698	.720
COP2_E	=	=	.686	.709
COPEN_E	=	=	.698	.720
COP1_IL	=	=	.693	.716
COP2_IL	=	=	.688	.710
COPEN_IL	=	=	.693	.716
COPS_E	=	=	.694	.716
COS_IL	=	=	.704	.726
COPS_E$^+$	.	=	.694	.717
COS_IL$^+$	=	.	.696	.719

Notes. The symbols $<$, $\ll$ and $\lll$ indicate that model j (column index) surpasses model i (row index) at confidence levels of 90%, 95% and 99%, respectively. Failure to reject the null hypothesis of equal predictive performance at a 10% significance level is indicated with "=".

A2.8 (Pseudo) out-of-sample time-series exercise for CFNAI

In the prediction exercise described in Section 2.3.4, we randomly split the sample into 3 independent sub-samples: training, testing and a fold of 21 observations to estimate the combinations of forecasts. Here, we test the robustness of the proposed methods to time-series frameworks, we re-conducted the forecasting exercise in CFNAI using an

expanding window to estimate the individual models and to form the combinations of forecasts.

We start our prediction exercise in December 2014. The training phase uses data available up to time t and then we extends the data by using an expanding window. The training set, made of the data available up to time t, is partitioned in two folds. The combination fold is composed of 21 observations selected randomly from the training set, and is used to estimate the combining weights. The other fold is used to train the individual models. Notice that the first iteration of this expanding window setup and the exercise described in Section 2.3.4 count the same number of observations available for training (146 for the estimation of the candidate forecasts and 21 to form forecast combinations). However, in the time series exercise, the size of the training set increases by one observation at each iteration, expanding the available information to estimate the baseline forecasts (e.g., in the last iteration, we count 187 observations for the estimation of the candidate forecasts and 21 to form the forecast combinations).

Because of the time-series dimension, we compute the Diebold-Mariano statistics by considering Newey-West standard errors (Newey and West, 1987) where we select the bandwidth of the Bartlett kernel using an AR(1) process. In the MCS test, we instead consider a block-bootstrap procedure with window 4 observations (4 months). In the MCS test, because of the time-series dimension, we consider a block-bootstrap procedure with window of 4 observations window.

Table A2.4: Predicting Chicago Fed National Activity Index (CFNAI), (pseudo) out-of-sample performance in time-series.

	Diebold-Mariano		RMSFE	MCS
i \ j:	COPS-E$^+$	COS-IL$^+$		
OLS	$\lll$	$\lll$	0.928	
L1	=	=	0.512	
L2	=	=	0.512	
svmLASSO	$\ll$	$\ll$	0.524	
svmRIDGE	$\lll$	$\lll$	0.556	
EN	$<$	$<$	0.517	
PLS	$<$	$<$	0.535	
RF	=	=	0.477	*
GBRT	=	=	0.513	
NN1	$<$	$<$	0.534	
NN2	$<$	$<$	0.531	
NN3	$\ll$	$\ll$	0.572	
NN4	$<$	$<$	0.549	
SA	=	=	0.481	*
TSA_3	=	=	0.476	*
TSA_5	=	=	0.486	
TSA_7	=	=	0.488	
TSA_9	=	=	0.490	
TSA_11	=	=	0.504	
T-MCS-60%	=	=	0.490	*
T-MCS-80%	=	=	0.494	
T-MCS-90%	=	=	0.525	
IL	=	=	0.478	*
DS_egal_lasso	$\lll$	$\lll$	0.557	
CO	$\lll$	$\lll$	0.657	
COPEN	=	=	0.461	*
CO$^+$	$\lll$	$\lll$	0.505	
COP1_E$^+$	=	=	0.463	*
COP2_E$^+$	=	=	0.496	
COPEN_E$^+$	=	=	0.467	*
COPE_E$^+$	=	=	0.460	*
COP1_IL$^+$	=	=	0.465	*
COP2_IL$^+$	=	=	0.500	
COPEN_IL$^+$	=	=	0.471	*
COPE_IL$^+$	=	=	0.465	*
COP1_E	=	=	0.467	*
COP2_E	=	=	0.498	
COPEN_E	=	=	0.494	
COP1_IL	=	=	0.501	
COP2_IL	=	=	0.495	
COPEN_IL	=	=	0.485	*
COPS_E	$<$	$<$	0.527	
COS_IL	$<$	$<$	0.540	
COPS_E$^+$	·	=	0.484	*
COS_IL$^+$	=	·	0.483	*

Notes. The symbols $<$, $\ll$ and $\lll$ indicate that model j (column index) surpasses model i (row index) at confidence levels of 90%, 95% and 99%, respectively. Failure to reject the null hypothesis of equal predictive performance at a 10% significance level is indicated with "=". The symbol $*$ marks the strategies that join the superior set of models at 10% significance level.

Chapter 3

Meta-learning approaches for recovery rate prediction

with Paolo Gambetti and Frédéric Vrins

Abstract

While previous academic research highlights the potential of machine learning and big data for predicting corporate bond recovery rates, the operations management challenge is to identify the relevant predictive variables and the appropriate model. We use *meta-learning* to combine the predictions from 20 candidate linear, nonlinear, and rule-based algorithms and we exploit a data set of predictors including security-specific factors, macro-financial indicators, and measures of economic uncertainty. We find that the most promising approach consists of model combinations trained on security-specific characteristics and a limited number of well-identified theoretically sound recovery rate determinants, including uncertainty measures. Our research provides useful indications for practitioners and regulators targeting more reliable risk measures in designing micro and macro-prudential policies.

Introduction

Credit risk is the main concern for financial institutions. It is governed by three factors: the exposure at default (EAD), the default likelihood (that is, the probability of default, PD), and the loss given default (LGD). The latter can also be expressed as $(1 - RR)$, where the recovery rate RR represents the percentage of EAD that can be recovered in the event of default of the reference entity. For a long time, financial institutions focused on the modeling of PDs. Ratings, for example, essentially quantify the risk of a default event, with the loss severity receiving little attention. However, EAD and RR are key drivers of credit risk. For instance, they act as scaling constants when computing banks' regulatory capital requirements (Loterman, Brown, Martens, Mues, and Baesens, 2012; Zhang and Thomas, 2012; BCBS, 2017) in Basel II (BCBS, 2006) and Basel III frameworks (BCBS, 2011), that banks can estimate using internal models in the advanced (A-IRB) approach. But RRs also influence the value of non-performing loans (Bellotti, Brigo, Gambetti, and Vrins, 2021), of corporate bonds, credit derivatives (Pykthin, 2003; Andersen and Sidenius, 2004; Berd, 2005; Gambetti, Gauthier, and Vrins, 2018), as well as the price of mainstream OTC derivatives products such as equity options and interest rates swaps due to counterparty risk (Gregory, 2012). This explains why recovery rates modeling and forecasting receive more and more attention nowadays.

Academic research highlights how nonlinear techniques (Qi and Zhao, 2011; Loterman, Brown, Martens, Mues, and Baesens, 2012; Tobback, Martens, Gestel, and Baesens, 2014;

Yao, Crook, and Andreeva, 2015; Nazemi, Heidenreich, and Fabozzi, 2018; Nazemi and Fabozzi, 2018) and ensemble methods (Bastos, 2014; Hartmann-Wendels, Miller, and Töws, 2014; Nazemi, Pour, Heidenreich, and Fabozzi, 2017; Bellotti, Brigo, Gambetti, and Vrins, 2021) should be preferred to the traditional parametric regressions used in earlier studies on recovery rate determinants. This finding is not only confined to *RR* forecasting but also in credit scoring (Wang, Liu, and Qi, 2022; Liu, Zhang, and Fan, 2022). However, the possibility of combining predictions from a large number of different algorithms remains widely unexplored. This is surprising given that it is now well recognized that combining models might be preferable (Atiya, 2020) to select a single model. In this chapter we confirm this conclusion by showing that meta-learning techniques exhibit better forecasting performance and lower model risk than methods in previous research when dealing with *RR*.[1]

Meta-learning can be defined as the act of a) selecting and b) optimally combining the outputs of multiple learning machines (*first-level learners*), according to some combination scheme that is learned from the data (*meta-learner or second-level learner*) (Santos, d. Araujo, dos Santos, Schwartz, and Menotti, 2017). Meta-learning is very flexible: the models to be combined can be represented by expert judgment, individual learners (e.g., Lasso, MARS), or *ensemble* learners (e.g., random forests, boosted trees), whose predictions themselves are already combinations of the outputs of different models. Furthermore, the models to combine can also be trained with different predictors. This contrasts with homogeneous ensemble methods in which, despite model combinations are involved, they generally combine weak learners[2] of the same type, e.g., regression trees are combined into a random forest.

The main advantage of meta-learning is hence the ability to exploit a wide spectrum of functional forms. Its promising applicability for recovery rate prediction was first reported in Nazemi, Pour, Heidenreich, and Fabozzi (2017) using a fuzzy decision fusion approach. However, several other alternatives are worth exploring. For example, Roccazzella, Gambetti, and Vrins (2022b) propose a robust and sparse combination method that mitigates estimation uncertainty and implicitly features forecast selection in a single step. This approach has the additional advantage of relying on a closed-form solution for the estimation of the optimal combination strategy.

The quality of the forecasts depends also on the predictive variables we consider. Prior studies on defaulted bonds highlight the importance of security-specific characteristics (Altman and Kishore, 1996; Schuermann, 2004; Bris, Welch, and Zhu, 2006; Jankowitsch, Nagler, and Subrahmanyam, 2014), economic conditions and the credit cycle (Altman, Brady, Resti, and Sironi, 2005; Acharya, Bharath, and Srinivasan, 2007; Bruche and González-Aguado, 2010; Betz, Kellner, and Rösch, 2018; Betz, Krüger, Kellner, and Rösch, 2020).[3] Models based on big data and variable selection techniques have been

[1] We define model risk from three perspectives: i) maximum average loss across model specifications and model classes, ii) average loss and iii) its variability within each model class.

[2] A weak learner is any machine learning algorithm that provides an accuracy slightly better than random guessing.

[3] Models based on economic principles approximate the latter using industry default rates, loan delinquency rates, market and industrial production returns and recession indicators (Altman, Brady, Resti, and Sironi, 2005; Jankowitsch, Nagler, and Subrahmanyam, 2014; Mora, 2015; Gambetti, Gauthier, and Vrins,

proposed by Nazemi, Pour, Heidenreich, and Fabozzi (2017); Nazemi, Heidenreich, and Fabozzi (2018) and Nazemi and Fabozzi (2018). Gambetti, Gauthier, and Vrins (2019) show that economic uncertainty is the most important systematic determinant of recovery rate distributions. However, the latter study is based on an ex-post analysis. Given that uncertainty shapes the economic outlook (ECB, 2009; Kose and Terrones, 2012; ECB, 2016; Gieseck and Largent, 2016) and its proxies are particularly capable of anticipating economic downturns (Ludvigson, Ma, and Ng, 2015), it remains to be explored whether uncertainty proxies can also improve predictive performance.

In this chapter, we extend the literature on recovery rate modeling studying the forecasting performance of linear, nonlinear rule-based, and meta-learning algorithms across different specifications of the predictors set.

We start our analysis by studying whether a larger set of macro-financial indicators, uncertainty measures, and idiosyncratic features in bond recovery rate forecasting models offers significant advantages compared to more parsimonious but economically grounded frameworks. Specifically, we extend the set of macroeconomic predictors used in Nazemi, Pour, Heidenreich, and Fabozzi (2017) and Nazemi, Heidenreich, and Fabozzi (2018) with 55 pricing factors and industry portfolio returns. We also enlarge the spectrum of uncertainty proxies considered by Gambetti, Gauthier, and Vrins (2019) with 11 additional measures of economic uncertainty derived from text-analysis techniques. This offers two advantages. First, as Jurado, Ludvigson, and Ng (2015a) point out, text-based indexes can show significant independent variations from other popular uncertainty proxies, suggesting that much of the variation in these proxies is not driven by uncertainty itself. Second, textual data is more granular, allowing us to identify the *categories* of uncertainty, e.g. government spending uncertainty or monetary policy uncertainty, that better predict recovery rates. In contrast to previous studies (Nazemi, Pour, Heidenreich, and Fabozzi, 2017; Nazemi, Heidenreich, and Fabozzi, 2018; Nazemi and Fabozzi, 2018), we find that more parsimonious models that rely on well-documented recovery rate determinants outperform those making use of the entire set of predictors and those relying on data-driven variable selection. A limited number of economically grounded predictors also makes the model easier to implement and manage. Among those, uncertainty measures are particularly relevant for recovery rate prediction. Moreover, it makes its results more transparent, hence, more prone to regulatory validation. This is consistent with the regulatory requirements of IRB approaches (BCBS, 2006, 2017).

Second, we provide the largest benchmark study of machine learning methods in the context of bond recovery rate prediction. We consider a total of 20 predictive algorithms and we obtain similar conclusions to those of Bellotti, Brigo, Gambetti, and Vrins (2021) in the case of recovery rates for nonperforming loans: random forests, quantile random forests, boosted trees, and cubist display the most promising performance seem to be the best class of models. Bagged MARS and model-averaged neural networks also seem to be competitive.

Third, we empirically show that meta-learning can be used to improve recovery rate predictions compared to traditional machine learning methods while considerably reducing model risk. This evidence is preserved both when looking at the predictive performance within a chosen predictor set and when considering joint predictions across

2019).

all specifications of the predictor set. The proposed methods outperform competitive combining methods such as the equally weighted forecast and the Hill Climbing algorithm of Caruana, Niculescu-Mizil, Crew, and Ksikes (2004), which showed promising results in the field of credit risk scoring (for more details Lessmann, Baesens, Seow, and Thomas, 2015). Furthermore, despite the main concern of this chapter is recovery rate predictions, the idea of combining heterogeneous models trained using diverse predictor sets is not specific to credit risk modeling and it can be used to hedge model uncertainty across the whole field of management science.

The remainder of the chapter is structured as follows. Section 3.1 contains an overview of the machine learning algorithms involved in our meta-learning approach and explains the latter. Section 3.2 provides a description of the data. Section 3.3 describes the main results of our benchmark study. Finally, we discuss the practical implications implied by our results and conclude the chapter.

3.1 Methodology

An overview of our predictive strategy can be found in Figure 3.1. After specifying the bond recovery rate predictors, first-level learners are trained to minimize the mean square error (MSE). Subsequently, the fitted residuals of the first-level learners (*meta-data*) become the input of second-level learners (*meta-learners*), which combine the original models with the goal of making the aggregate forecast error variance as small as possible. Finally, we evaluate the predictive performance of the various classes of linear, nonlinear, rule-based, and meta-learning methods.

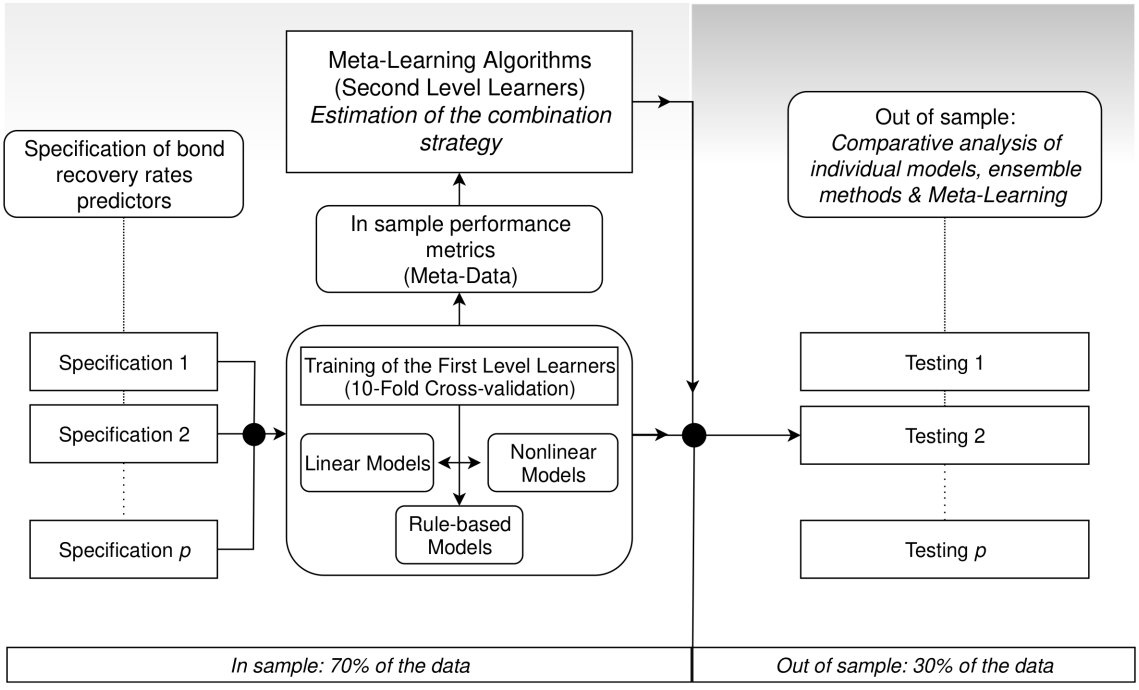


Figure 3.1: Predictive strategy with meta-learning techniques.

We briefly review the predictive algorithms and combination strategies that will serve first- and second-level learners, respectively.

3.1.1 First-level learners

To undertake an unbiased benchmark study, the spectrum of competing models should be as rich as possible. This requirement has rarely been respected in the context of recovery rate modeling, except for the large-scale benchmark studies of Loterman, Brown, Martens, Mues, and Baesens (2012) and Bellotti, Brigo, Gambetti, and Vrins (2021).

For comparison purposes, we use a similar set of techniques as Bellotti, Brigo, Gambetti, and Vrins (2021), which employ 20 algorithms belonging to three different classes: linear, nonlinear, and rule-based.

We provide the list in Table 3.1 together with the corresponding `R` implementations. For more details on nonlinear and rule-based methods, we refer to Appendix A. [4]

Table 3.1: List of first-level learners and corresponding `R` algorithms. The three blocks of the table identify linear (top), nonlinear (middle), and rule-based (bottom) models.

Description	Acronym	`R` algorithm	Reference
Linear regression	lm	`lm`	R Core Team (2017)
Backward step-wise selection	lm_bs	`leaps`	Lumley (2017)
Ridge regression[(a)]	ridge	`glmnet`	Friedman, Hastie, Tibshirani, Balasubramanian, and Noah (2019)
Lasso regression[(a)]	Lasso	`glmnet`	"
Elastic net regression	elnet	`glmnet`	"
MARS	mars	`earth`	Milborrow (2018)
Bagged MARS	bmars	`earth`	"
Model averaged neural networks	avnnet	`nnet`	Ripley and Venables (2016)
Support vector regression	svr	`ksvr`	Karatzoglou, Smola, Hornik, and Zeileis (2004)
Relevance vector regression	rvm	`rvm`	"
Gaussian processes	gauss	`gausspr`	"
Regression trees	cart	`rpart`	Therneau, Atkinson, and Ripley (2017)
Conditional inference trees	cit	`ctree`	Hothorn, Hornik, and Zeileis (2006)
Boosted tree	bst	`bst`	Wang (2018)
Stochastic gradient boosting	gbm	`gbm`	Greenwell, Boehmke, Cunningham, and Developers (2018)
Random forests	rf	`randomForest`	Liaw and Wiener (2002)
Quantile random forests	qrf	`quantregForest`	Meinshausen (2017)
Cubist	cubist	`cubist`	Kuhn and Quinlan (2018)

Notes. (a) We consider two versions of the ridge and Lasso: i) the standard one involving the penalty term associated with the best-in-sample performance and ii) that based on the one-standard-error rule of Hastie, Tibshirani, and Friedman (2009).

3.1.1.1 Linear models

We consider seven linear models with and without penalization. Following Bellotti, Brigo, Gambetti, and Vrins (2021), they can be framed as the following minimization problem:

$$\underset{\beta_0,\, \boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \beta_0\|_2^2 + \lambda \left((1-\alpha)\|\boldsymbol{\beta}\|_2^2 + \alpha\|\boldsymbol{\beta}\|_1\right) \tag{3.1}$$

where $\boldsymbol{y}$ denotes the vector of the observations in the sample, $\boldsymbol{X}$ is the N-by-p matrix of predictors and $\boldsymbol{\beta}$ stands for the vector of regression coefficients. Different specifications of the penalty term $\lambda \geq 0$ and the mixing factor $0 \leq \alpha \leq 1$ yield the standard linear regression model (with and without backward selection), ridge, Lasso, and elastic net. Models of these types can only reproduce linear relationships such as those reproduced in Figure 3.2.

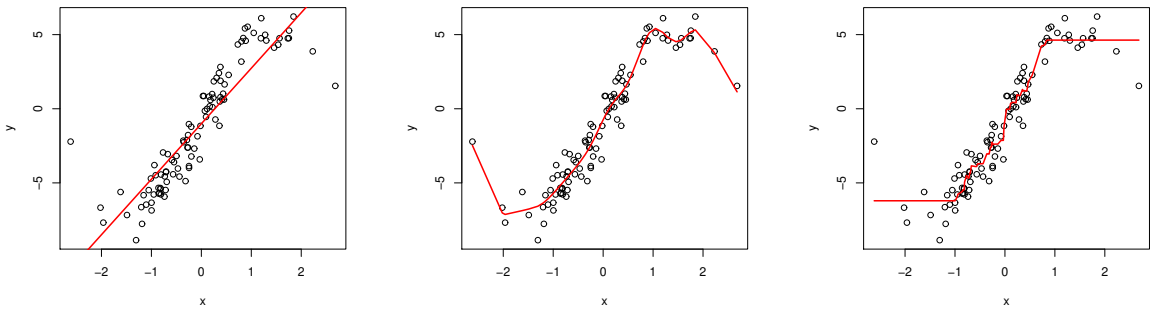
[4] Hyperparameters for first-level learners are tuned using 10-fold cross-validation in the training sample. Folds are created using stratified sampling based on seniority type, as in Nazemi, Pour, Heidenreich, and Fabozzi (2017), Nazemi and Fabozzi (2018) and Nazemi, Heidenreich, and Fabozzi (2018). The same applies to generating the training and test sets, with proportions of 70% and 30%.

3.1.1.2 Nonlinear models

We deal with six types of nonlinear models. Among the kernel methods, we consider support vector regression (SVR), relevance vector machines (RVM) regression, and Gaussian processes. We further consider multivariate adaptive regression splines (MARS) and two nonlinear ensembles: model-averaged neural networks and bagged MARS. We refer the reader to Bellotti, Brigo, Gambetti, and Vrins (2021) for an extended treatment of each of them. These models are naturally suited to capture nonlinear relationships, but they can be prone to overfit. An example of nonlinear fit is included in Figure 3.2.

3.1.1.3 Rule-based models

According to Gambetti, Gauthier, and Vrins (2019), rule-based methods are able to identify clusters of data with similar properties and reproduce step-like relationships (with or without slopes depending on the particular algorithm). We include seven types of rule-based methods. Individual models include regression trees and conditional inference trees, while ensembles are represented by cubist, random forests, quantile random forests, and boosted trees with and without stochastic gradient boosting. An example of rule-based model fit is included in Figure 3.2.



(a) Linear model. **(b)** Nonlinear model. **(c)** Rule-based model.

Figure 3.2: Examples of fitted relationships for linear, nonlinear and rule-based models. The true data generating process (dots) follows a shifted sine wave with Gaussian noise. Panel (a) represents the fit obtained by linear regression. Panel (b) represents the fit obtained by support vector regression. Panel (c) represents the fit obtained by boosted trees with stochastic gradient boosting.

3.1.2 Second-level learners

Starting from m predictions $\widehat{y}_1, \ldots, \widehat{y}_m$ of the random variable y returned by m first-level learners, two strategies can be adopted. The standard procedure consists of selecting the best forecast, say $\widehat{y} := \widehat{y}_{i^\star}$, where model $i^\star$ is selected according to some criteria. Alternatively, these predictions can be aggregated to form a single prediction $\widehat{y} := \phi(\widehat{y}_1, \ldots, \widehat{y}_m)$ using some function ϕ. The model combination offers diversification gains that make it attractive when we cannot identify ex-ante the best single model.[5] In addition, in the rare cases where the best model can be identified, meta-learning

[5]Forecast selection outperforms the forecast combination only in very specific situations that are typically not encountered in practice: for instance, when the variance of the prediction errors of one model is

techniques could still take full advantage of the available information when the first-level learners rely on various data sources or cover a wide spectrum of modeling assumptions.

Meta-learning learns the combination strategy ϕ directly from the data with the explicit goal of minimizing a loss function. Precisely, let $\mathbf{y}$ be an n-dimensional column vector containing n observations of the target variable and $\widehat{\mathbf{Y}}$ be an n-by-m matrix of m unbiased candidate forecasts of $\mathbf{y}$. The optimal combination strategy consists of estimating the function $\phi(\widehat{\mathbf{Y}})$ that solves

$$\phi^* := \underset{\phi \in \Phi}{\arg\min} \left\| \boldsymbol{y} - \phi(\widehat{\boldsymbol{Y}}) \right\|_2^2 , \tag{3.2}$$

where $\Phi := \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ is some conformable functional space.

Nevertheless, its success depends on how accurately the combination strategy can be determined. The use of in-sample data both to train the individual models and to consecutively combine them on the basis of their respective fitted residuals can significantly amplify the initial estimation error and consequently produce poor out-of-sample predictions.

Meta-learning relies on validation techniques to assess how well the combination weights will generalize to the out-of-sample predictive exercise. We opt to combine by minimizing the variance of the aggregate prediction error and include robust combination methods to hedge against potential instability[6].

3.1.2.1 Linear meta-learners

Restricting ourselves to the class of linear combinations with weight summing to one, i.e., for Φ being the set of functions from $\mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ of the form $\phi(x_1, \ldots, x_m) := \sum_{i=1}^m w_i x_i$ satisfying $\sum_{i=1}^m w_i = 1$ leads to a combined prediction of the form $\widehat{y} = \sum_{i=1}^m w_i \widehat{y}_i$. In this case, the optimization problem becomes

$$\boldsymbol{w}^* := \underset{\boldsymbol{w} \in \mathscr{W}}{\arg\min} \left\| \boldsymbol{y} - \widehat{\boldsymbol{Y}} \boldsymbol{w} \right\|_2^2 \quad \text{where} \quad \mathscr{W} := \{ \boldsymbol{w} \in \mathbb{R}^m \mid \mathbf{1}^T \boldsymbol{w} \ = \ 1 \} . \tag{3.3}$$

Constraining the weights to sum to 1 keeps the aggregate prediction unbiased provided that so are the candidate forecasts (which is the case in this chapter). By additionally constraining the weights to be nonnegative, this approach corresponds to Breiman's stacked regressions (Breiman, 1996b), where linear combinations of different predictors are considered to further improve prediction accuracy. Granger and Ramanathan (1984) show that this approach is equivalent to selecting the weights $\boldsymbol{w}^*$ that minimize the variance of the meta-learner's prediction error.

lower than those of the others by several orders of magnitude, see, e.g., Roccazzella, Gambetti, and Vrins (2022b).

[6] Another strategy consists of using an additional validation fold (Wolpert, 1992). This has the drawback of extending the original data with potentially informative observations that would unevenly boost the performance of meta-learning techniques with respect to those of individual models and ensemble methods. In this chapter, we empirically show that combining schemes that do not rely on additional sample splitting perform remarkably well compared to the ex-post best predictive framework. This is surprising, especially for combination schemes whose weights are estimated using the same in sample information

Nevertheless, this combination strategy can lack robustness when the covariance matrix of prediction errors is poorly estimated. This often occurs because of sample size limitations, considerable background noise, or when first-level forecasts are highly collinear (Claeskens, Magnus, Vasnev, and Wang, 2016). We can tackle the potential instability of $\boldsymbol{w}^*$ by using the linear shrinkage estimator of the covariance matrix of prediction errors, which we denote by $\boldsymbol{\Sigma}_\lambda$. The latter consists of combining the sample covariance matrix (which is easy to compute, asymptotically unbiased but prone to estimation errors) with an estimator that is misspecified and biased but more robust to estimation errors (Ledoit and Wolf, 2004a). This approach, initially derived in a portfolio optimization context, was recently transposed in Roccazzella, Gambetti, and Vrins (2022b) to the forecast combination problem, showing that constrained optimization with shrinkage (COS) of $\boldsymbol{\Sigma}$ can provide a single-step, fast and robust optimal forecast combination strategy. Here we adapt the COS to act as a linear meta-learner, which combines the first-level learners only on the basis of in-sample information. This leads to the following optimization problem:

$$\boldsymbol{w}^*_\lambda := \underset{\boldsymbol{w} \in \mathscr{W}^+}{\arg\min}\ \boldsymbol{w}^T \boldsymbol{\Sigma}_\lambda \boldsymbol{w} \quad \text{where} \quad \boldsymbol{\Sigma}_\lambda := (1-\lambda)\,\boldsymbol{\Sigma} + \lambda \overline{\boldsymbol{\Sigma}}\,, \tag{3.4}$$

$\lambda \in [0,1]$ is the shrinkage intensity, $\boldsymbol{\Sigma}$ is the *sample* covariance matrix of prediction errors and $\overline{\boldsymbol{\Sigma}}$ is a predetermined reference covariance matrix. We estimate the optimal shrinkage intensity by minimizing the expected Frobenius norm of the difference between $\boldsymbol{\Sigma}_\lambda$ and the *population* covariance matrix of prediction errors $\boldsymbol{S}$ (Ledoit and Wolf, 2004a).[7]

$$\lambda^* := \underset{\lambda \in [0,1]}{\arg\min}\ \mathbb{E}\,\|\boldsymbol{\Sigma}_\lambda - \boldsymbol{S}\|_2^2\ . \tag{3.5}$$

We consider two shrinkage directions for $\overline{\boldsymbol{\Sigma}}$.

Definition 3.1.1 (COS-E -*Constrained Optimization with Shrinkage towards Equal weights*)**.** *The target covariance matrix $\overline{\boldsymbol{\Sigma}}_{COS-E}$ corresponds to the case where first-level learners are assumed to have identical prediction error variance $\bar{\sigma}^2$ and identical pairwise correlation coefficients $\bar{\rho}$:*

$$\overline{\boldsymbol{\Sigma}}_{COS-E} := \overline{\sigma}^2 \begin{bmatrix} 1 & \overline{\rho} & \dots & \overline{\rho} \\ \overline{\rho} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \overline{\rho} \\ \overline{\rho} & \dots & \overline{\rho} & 1 \end{bmatrix}\ . \tag{3.6}$$

In practice, we set $\bar{\sigma}^2 = \frac{1}{m}\sum_{i=1}^m \sigma_i^2$ where σ_i^2 is the sample prediction error variance of model i (the average error prediction variance in the set of first-level learners) and by $\bar{\rho} = \frac{2}{m(m-1)}\sum_{i=1}^{m-1}\sum_{j=i+1}^{m}\rho_{i,j}$ (the average correlation coefficient of first learners' in-sample prediction error).

Definition 3.1.2 (COS-IL -*Constrained Optimization with Shrinkage towards Inverse Loss-based weights*)**.** *The target covariance matrix $\overline{\boldsymbol{\Sigma}}_{COS-IL}$ corresponds to the case where*

[7] For further details on the COS methodology and for the explicit formula to estimate the optimal shrinkage intensity, we refer to Roccazzella, Gambetti, and Vrins (2022b).

first-level learners have an identical pairwise correlation coefficient $\bar{\rho}$, but their respective prediction error variance is estimated using sample data:

$$\overline{\mathbf{\Sigma}}_{COS-IL} := \begin{bmatrix} \sigma_1^2 & \overline{\rho}\sigma_1\sigma_2 & \dots & \overline{\rho}\sigma_1\sigma_m \\ \overline{\rho}\sigma_2\sigma_1 & \sigma_2^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \sigma_{m-1}^2 & \overline{\rho}\sigma_{m-1}\sigma_m \\ \overline{\rho}\sigma_m\sigma_1 & \dots & \overline{\rho}\sigma_m\sigma_{m-1} & \sigma_m^2 \end{bmatrix}. \tag{3.7}$$

As benchmarks, we also include the equally weighted forecast, i.e., $\phi(x_1, \dots, x_m) := \sum_{i=1}^{m} \frac{1}{m} x_i$ and the Hill Climbing algorithm of Caruana, Niculescu-Mizil, Crew, and Ksikes (2004). In this latter case, forward step-wise selection is used to identify the equally weighted combination that minimizes the Root Mean Square Error RMSE in a validation fold sampled from the training data.

3.1.2.2 Nonlinear meta-learners

Thus far, we have described linear weighting schemes. Nevertheless, meta-learning is more general. For example, in the field of image classification, meta-learning techniques typically involve nonlinear methods such as deep learning or recurrent models (Santoro, Bartunov, Botvinick, Wierstra, and Lillicrap, 2016), metric learning (Koch, 2015), and learning optimizers (Ravi and Larochelle, 2017). Despite being more flexible than linear methods, nonlinear meta-learners are also more complex and prone to overfitting, especially in relatively small data set.[8] For these reasons, we opt for ensembles of shallow (one hidden layer) feed-forward neural networks (NNs), which can still approximate any measurable function at any desired degree of accuracy provided that sufficiently many hidden units are available (Hornik, Stinchcombe, and White, 1989). In this case, we consider Φ to be the set of functions from $\mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^k$ of the form $\phi(x_1, \dots, x_m) := \sum_{i=1}^{k} \frac{w_i}{1+\exp\{-\sum_{j=1}^{m} \alpha_{ij} x_j\}}$. where k is the number of neurons in the hidden layer.

As for the linear combination schemes, we train the NNs to minimize the in-sample MSE using as input the first-level predictions. However, in this case, the output of the meta-learning is not a linear function of the inputs, and the combining weights are not constrained to sum to one. Therefore, studying the marginal contribution of each first-level prediction to the aggregate prediction is not straightforward. In our empirical exercise, we consider two feed-forward neural networks with 1 (NN-1) and 2 (NN-2) hidden layers and 8 and 16-8 neurons, respectively.

3.2 Data

3.2.1 Recovery rates and security-specific characteristics

The main references on bond recovery rate modeling are generally based on Moody's Default & Recovery Database (Moody's DRD) or the Standard & Poor's Capital IQ Database. We employ Moody's DRD in this chapter.

[8] For example, Dodge and Karam (2016) documents that deep learning methods are particularly sensitive to noise levels in image classification tasks.

Following the standard market convention (Schuermann, 2004; Mora, 2015), the recovery rate of each bond is computed as the bond price measured 30 days after the default date, declared by the rating agency, and divided by the face value. To make our analyses more comparable with earlier studies such as Nazemi, Pour, Heidenreich, and Fabozzi (2017), Nazemi, Heidenreich, and Fabozzi (2018), and Nazemi and Fabozzi (2018), we apply a similar filtering strategy. We selected dollar-denominated bonds issued by U.S. companies and with at least USD 5 million in face value. To replicate the same economic conditions, we also filter for default issues in the period 2002-2012. We only retain the observations associated with known values for the following security-specific characteristics: debt seniority, issuer's industrial sector, default type, coupon level, maturity, presence of additional guarantees different from the issuer's asset, and default date. We are thus left with 768 observations. Notice that machine learning methods provide useful insights even with a sample of this size. In particular, the considered predictive models mitigate the risk of over-fitting, while identifying the most relevant predictive variables and exploring the existence of potential nonlinearities in the data (nonlinear and rule-based methods).

Figure 3.3 includes a histogram of our recovery rate sample. We provide the corresponding summary statistics in Table 3.2.

Summary statistics of recovery rates conditional on the seniority of the defaulted bond, issuer's industrial sector, and default type are provided in Tables 3.3, 3.4 and 3.5, respectively. They are consistent with previous findings on recovery rate determinants. For instance, bonds with higher seniority are associated with higher recovery rates on average, and those of senior secured bonds are the most dispersed (Altman and Kishore, 1996; Altman and Kalotay, 2014). Recovery rates are also higher, on average, when the issuer is operating in an industrial sector featuring higher asset tangibility and in the utility sector in particular (Altman and Kishore, 1996; Schuermann, 2004; Acharya, Bharath, and Srinivasan, 2007). Similarly, milder default procedures lead to higher recovery rates (Franks and Torous, 1994; Bris, Welch, and Zhu, 2006; Davydenko and Franks, 2008; Altman and Karlin, 2009). Defaults on security cash flows are expected to recover more than company reorganizations or liquidations. Controlled reorganizations (prepackaged Chapter 11) also display higher recovery than uncontrolled reorganizations and asset liquidations (receivership and procedures included in the "others" category). Tables 3.6, 3.7 and 3.8 show that our sample of recovery rates is also consistent with prior findings about the role of coupons, maturity, and backing guarantees. Recovery rates are higher in the presence of backing guarantees, increase with coupons and decrease with maturity (Jankowitsch, Nagler, and Subrahmanyam, 2014). All these security-specific characteristics have been used as control variables in recent articles on bond recovery rate determinants (Gambetti, Gauthier, and Vrins, 2019; François, 2019) and are used as predictors in our study.

Table 3.2: Summary statistics of our recovery rate sample.

	N	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	St.Dev.
Recovery rate	768	0.01%	10.00%	20.00%	30.98%	51.41%	118.00%	27.58%

Summary statistics of recovery rates conditional on the seniority of the defaulted bond, issuer's industrial sector, and default type are provided in Tables 3.3, 3.4 and 3.5, respectively. They are consistent with previous findings on recovery rate determinants.

For instance, bonds with higher seniority are associated with higher recovery rates on average, and those of senior secured bonds are the most dispersed (Altman and Kishore, 1996; Altman and Kalotay, 2014).

Table 3.3: Summary statistics of recovery rates according to the seniority of the defaulted bond.

Debt seniority	N	Median	Mean	St.Dev.	Skew.
Senior Secured	85	63.00%	60.92%	32.27%	−0.29
Senior Unsecured	533	19.00%	28.03%	23.90%	0.98
Senior Subordinated	129	19.13%	25.77%	26.87%	1.33
Subordinated	21	9.13%	16.71%	23.37%	2.20

Recovery rates are also higher, on average, when the issuer is operating in an industrial sector featuring higher asset tangibility and in the utility sector in particular (Altman and Kishore, 1996; Schuermann, 2004; Acharya, Bharath, and Srinivasan, 2007). Similarly, milder default procedures lead to higher recovery rates (Franks and Torous, 1994; Bris, Welch, and Zhu, 2006; Davydenko and Franks, 2008; Altman and Karlin, 2009). Defaults on security cash flows are expected to recover more than company reorganizations or liquidations. Controlled reorganizations (prepackaged Chapter 11) also display higher recovery than uncontrolled reorganizations and asset liquidations (receivership and procedures included in the "others" category). Tables 3.6, 3.7 and 3.8 show that our sample of recovery rates is also consistent with prior findings about the role of coupons, maturity, and backing guarantees. Recovery rates are higher in the presence of backing guarantees, increase with coupons and decrease with maturity (Jankowitsch, Nagler, and Subrahmanyam, 2014). All these security-specific characteristics have been used as control variables in recent articles on bond recovery rate determinants (Gambetti, Gauthier, and Vrins, 2019; François, 2019) and are used as predictors in our study.

Table 3.4: Summary statistics of recovery rates according to the industrial sector of the bond issuer.

Industrial sector	N	Median	Mean	St.Dev.	Skew.
Banking	18	18.00%	23.47%	24.44%	0.46
Capital Industries	189	29.00%	36.44%	28.55%	0.72
Consumer Industries	88	30.25%	39.86%	28.68%	0.49
Energy & Environment	45	40.00%	44.10%	25.76%	0.64
FIRE	166	10.00%	11.95%	10.32%	3.88
Media & Publishing	90	43.62%	40.72%	31.45%	0.001
Retail & Distribution	32	36.25%	34.92%	25.89%	0.58
Technology	72	15.00%	19.89%	16.99%	2.45
Transportation	57	22.25%	31.32%	23.81%	1.16
Utilities	11	92.50%	91.89%	6.54%	−0.32

Table 3.5: Summary statistics of recovery rates according to the default type.

Default type	N	Median	Mean	St.Dev.	Skew.
Chapter 11	371	10.50%	25.68%	25.97%	1.53
Missed interest payment	281	28.50%	37.55%	26.76%	0.53
Missed principal and interest payments	14	58.12%	56.14%	18.85%	0.05
Missed principal payment	8	23.04%	29.76%	29.12%	0.99
Others	6	11.00%	15.59%	17.47%	1.50
Prepackaged Chapter 11	71	12.00%	31.95%	33.01%	0.71
Receivership	7	0.50%	4.57%	9.70%	2.01
Suspension of payments	10	18.50%	30.05%	26.75%	2.01

Table 3.6: Average recovery rate by coupon level.

Coupon	$[0\% - 2.5\%)$	$[2.5\% - 5\%)$	$[5\% - 7.5\%)$	$[7.5\% - 10\%)$	$\geq 10\%$
Average RR	14.82%	24.22%	23.01%	33.11%	36.95%

Table 3.7: Average recovery rate by maturity level.

Maturity (years)	$[0 - 5\,\text{y})$	$[5\,\text{y} - 10\,\text{y})$	$[10\,\text{y} - 15\,\text{y})$	$[15\,\text{y} - 20\,\text{y})$	$\geq 20\,\text{y}$
Average RR	43.30%	37.88%	31.92%	19.39%	18.90%

Table 3.8: Average recovery rate for bonds with and without backing.

Backing	Yes	No
Average RR	40.02%	29.33%

3.2.2 Systematic factors

We extract industry default rates from Moody's DRD and the remaining systematic factors from the databases managed by the Federal Reserve Bank of St. Louis: FRED and FRED-MD (McCracken and Ng, 2015). The latter includes a large set of time series referring to output and income, labor market, housing, consumption, orders and inventories, money and credit, interest and exchange rates, prices, and the stock market [9].

We then extend this data set with a large number of economic uncertainty measures from three different classes: survey-based, news-based, and volatility-based.[10] All measures are retrieved from the websites of the authors (Baker, Bloom, and Davis, 2016; Jurado, Ludvigson, and Ng, 2015a; Ludvigson, Ma, and Ng, 2015). We consider all five measures employed in the study by Gambetti, Gauthier, and Vrins (2019) plus 11 additional measures of categorical economic policy uncertainty. Table 3.9 gives an overview. Factors

[9] We refer to the appendix for the full list of the variables and the transformations performed on the raw data.

[10] The reader can refer to Appendix B for more details and to Gambetti, Gauthier, and Vrins (2019) for a detailed literature review on the topic.

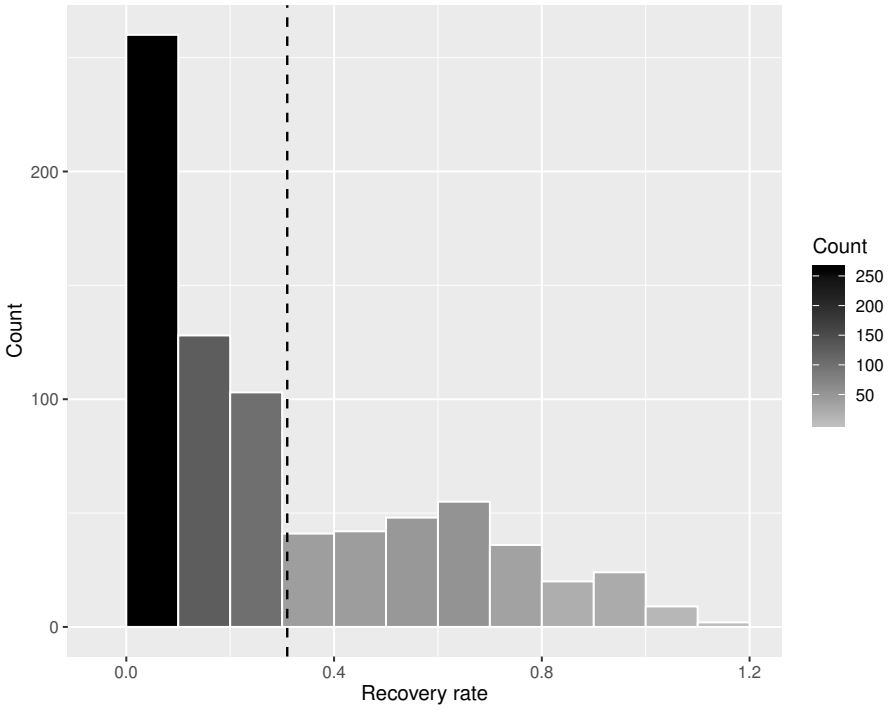


Figure 3.3: Histogram of recovery rates in our sample. The dashed line represents the sample mean. The average recovery rate is 30.98%, while the standard deviation is equal to 27.58, which is a large value compared to the mean. The distribution is also highly skewed, a typical feature of recovery rate data.

relating to the market price of risk and industry portfolio returns are instead retrieved from the Fama-French database. Predictors are measured one month before the default date.

Table 3.9: List of uncertainty measures considered in this study.

Name	**Type**	**Methodology**	**References**
Inflation un. Federal/State/Local expenditures un.	Survey	Dispersion of forecasts from the Federal Reserve Bank of Philadelphia's Survey of Professional Forecasters.	Zarnowitz and Lambros (1987) Bachmann, Elstner, and Sims (2013) Baker, Bloom, and Davis (2016)
Economic policy un. Monetary policy un. Fiscal policy (taxes or spending) un. Tax un. Government spending un. Health care un. National security un. Entitlement programs un. Regulation un. Financial regulation un. Trade policy un. Sovereign debt, currency crises un.	News	Normalized volume of newspaper articles published in a given month containing expressions referring to specific types of economic un..	Baker, Bloom, and Davis (2016) Alexopoulos and Cohen (2015)
VIX	Volatility	Stock market implied volatility index from the Chicago Board Options Exchange.	Bloom (2009) Bekaert, Hoerova, and Lo Duca (2013)
Financial un.	Volatility	Conditional volatility of the purely unforecastable prediction error of financial time-series.	Jurado, Ludvigson, and Ng (2015a) Ludvigson, Ma, and Ng (2015)

Notes. The abbreviation *un.* stays for *uncertainty*.

We now proceed to present the results and to discuss their relevance with respect to practical implementation in the industry.

3.3 Results

We compare the performance of the considered algorithms across nine specifications of the predictor set. All model specifications feature security-specific characteristics but differ in the systematic variables that are included. We use two approaches to determine the latter.

a) Statistical approach: the algorithms can access either the full data set of systematic variables or a set created with variable selection techniques. For variable selection, we consider a model based on Lasso-selected systematic variables and one based on Lasso with stability selection (Meinshausen and Bühlmann, 2010). While the latter has been used in Nazemi and Fabozzi (2018) to check the reliability of their Lasso-selected macroeconomic variables, those retained by Lasso with stability selection have never been used to feed predictive algorithms.[11]
b) Economic approach: we create models by relying exclusively on well-identified factors based on the results of Gambetti, Gauthier, and Vrins (2019) and prior studies on recovery rate determinants. Table 3.10 includes a summary of the model specifications.

Table 3.10: Description of the different model specifications.

Specification ID	Systematic variables	Reference
1	Full data set	-
2	Lasso-selected macroeconomic variables	Nazemi and Fabozzi (2018)
3	Lasso-selected variables with stability control	-
4	Industry default rate, commercial and industrial loans delinquency rates, industrial production, market index returns, PMI	$\mathscr{M}_{3+}$ in Gambetti, Gauthier, and Vrins (2019)
5	As in 4, with financial uncertainty substituted to industry default rates	$\mathscr{M}_{4+}$ in "
6	As in 4, with VIX substituted to industry default rates	$\mathscr{M}_{5+}$ in "
7	As in 4 plus Financial Uncertainty, news-based Economic Policy Uncertainty, Inflation uncertainty and Federal/State/local expenditures uncertainty	$\mathscr{M}_{4++}$ in "
8	As in 4 plus all uncertainty measures of Table 3.9	-
		$\mathscr{M}_2$ in "
9	No systematic variables	Nazemi and Fabozzi (2018) Nazemi, Heidenreich, and Fabozzi (2018)

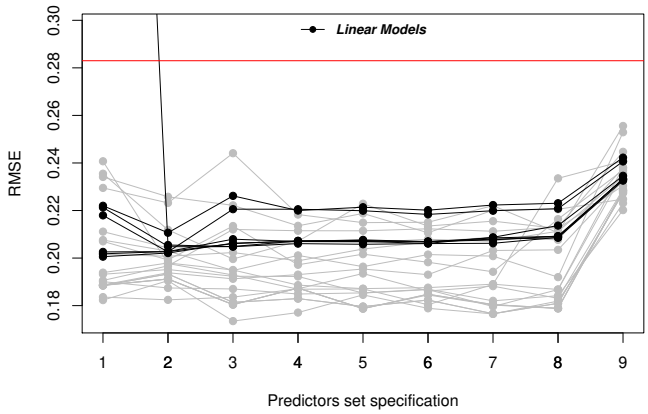
We compare the considered predictive strategies by analyzing the root mean square forecast error (RMFE) and we also assess the significance of a performance difference via a model confidence set test (hereafter MCS, Hansen, Lunde, and Nason, 2011). [12]

[11] We apply Lasso with stability selection based on the `R` implementation `stabs` by Hofner and Hothorn (2017). We determine the dimension of bootstrapped Lasso models using pointwise control (Meinshausen and Bühlmann, 2010). Moreover, we specify a threshold of .6 for the selection probability as in Nazemi and Fabozzi (2018).

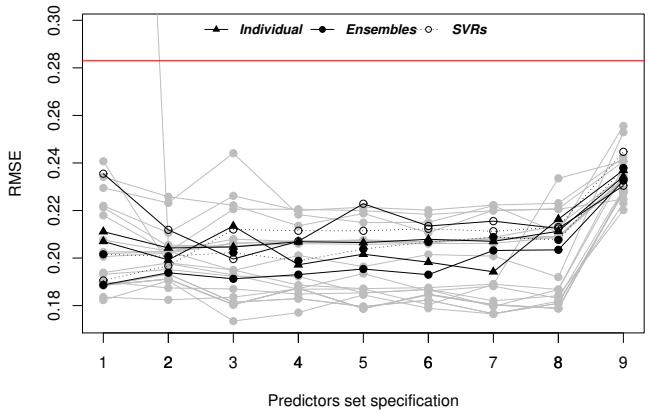
[12] The MCS tests whether a subset of methods enters jointly in the superior set of models by repeatedly testing the null hypothesis of equal predictive performance with significance level α. Let $\mathscr{M}_0$ be the set of all forecasting models (both individual candidates and forecast combinations), and let $\mathscr{M}^*$ be the superior set of models. Formally, the MCS tests $H_0 : \mathbb{E}[d_{i,j}] = 0, \ \forall i, j$. If the null hypothesis is rejected, then the procedure eliminates the model with the greatest relative loss from the set $\mathscr{M}$. This procedure is sequentially repeated until the null hypothesis is not rejected at the chosen probability level α. We compute the MCS p-values via bootstrapping (10000 replications) and using the Oxford MFE Toolbox publicly available at `https://www.kevinsheppard.com/code/matlab/mfe-toolbox/`.

3.3.1 Predictive models vs. historical averages

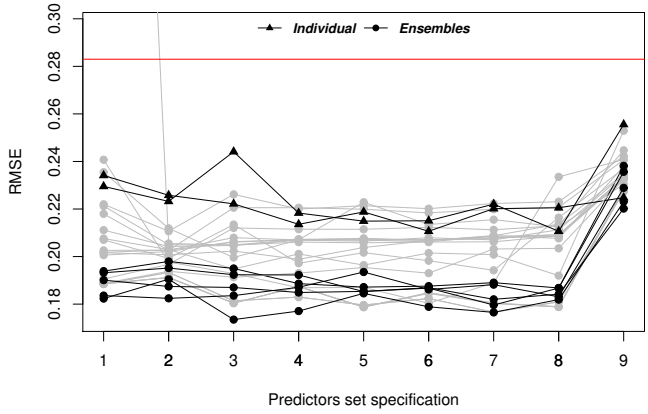
The performance of our predictive algorithms across model specifications is summarized in Figure 3.4; the detailed out-of-sample RMSE are reported in Table 3.12. Each line corresponds to a different algorithm. The solid horizontal line (in red) corresponds to the model where we use the in-sample mean recovery rate as a prediction; it can be assimilated to the regulatory standard approach and foundation-IRB Approach (BCBS, 2006, 2011), which are largely based on historical LGD values. Figure 3.4 highlights that forecasting recovery rates using an algorithm always leads to more precise estimates than using the average of previously observed values. The only exception is the linear regression, which, as expected, suffers from the curse of dimensionality when trained on the full set of predictors (specification 1).



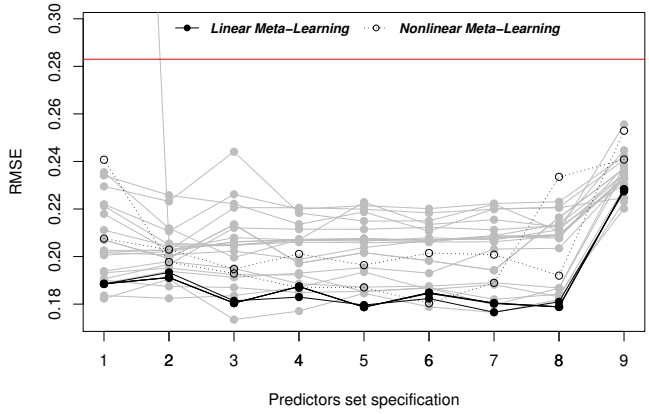
(a) Linear models.



(b) Nonlinear models.



(c) Rule-based models.



(d) Meta-learning

Figure 3.4: Illustration of forecasting performance across model specifications. The horizontal red line is the model based on the in-sample mean recovery rate. Panel (a) highlights the performance of linear models. Panel (b) highlights the performance of nonlinear models. Panel (c) highlights the performance of rule-based models. Panel (d) illustrates the performance of linear and non-linear meta-learning methods.

3.3.2 Models based on systematic variables

Table 3.12 and Figure 3.4 also make clear that, in line with the literature, models exploiting systematic variables (specifications 1 to 8) always yield better forecasting performance than a model based on security-specific factors alone (specification 9).

For the selection of macroeconomic variables with data-driven methods, we observe from Table 3.12 that Lasso selection seems to bring more benefit to linear models than to nonlinear or rule-based models. In particular, it appears that ensemble methods (neural networks, bagged MARS, boosted trees, random forests, quantile random forests, cubist) are deprived of useful predictive information when they are trained on Lasso-selected variables (specification 2). The performance of these latter algorithms improves if we instead implement the Lasso procedure with stability selection (specification 3). We obtain particularly competitive RMSE values in this latter case; the best performance of first-level learners reaches an RMSE of .174 with boosted trees. From an aggregate perspective, it is clear that models relying on Lasso selection (with or without stability control) reduce the embedded model risk compared to using the complete set of predictors.

We find that Lasso with stability selection retains financial uncertainty, uncertainty related to government spending and uncertainty related to sovereign debt and currency crises (Table 3.11) . Financial uncertainty is associated with the highest probability of being selected (95%). The second and third factors, an index of consumption expenditures and a housing market indicator, are associated with probabilities of 88% and 86%, respectively.[13] Given these results, we confirm the importance of uncertainty measures for forecasting bond recovery rates.

Moreover, if we consider the features of data-driven selection methods against those of economically motivated models, we should prioritize the use of the latter. Notwithstanding that both methods exhibit comparable predictive performance, the economically motivated models offer several advantages. By being based on a low-dimensional set of well-identified economic factors, they are easier to implement and monitor and yield more interpretable results. In contrast, data-driven selection methods can fail to correctly select the best subset of predictors when the latter feature high correlation and require the specification of additional hyperparameters. This increases the complexity and the computational burden, with no guarantee of perfectly selecting the full set of significant predictors.

3.3.3 First level learners

We now discuss the findings regarding the out-of-sample performance of our algorithms across model structures. In Figure 3.4, we highlight (in black) the performance of ensemble methods: model-averaged NNs, bagged MARS, boosted trees (with and without stochastic gradient boosting), cubist, random forests, and quantile random forests.

In this respect, our results point in the same direction as what Bellotti, Brigo, Gambetti, and Vrins (2021) discover for nonperforming loans. Rule-based ensemble methods are always associated with the most promising performance compared to both individual

[13] We find our list of selected variables to be largely consistent with those highlighted in Nazemi and Fabozzi (2018). A table of predictor probabilities is included in the appendix.

Table 3.11: List of predictor probabilities from Lasso with stability selection.

Variable	Selection probability
Financial Uncertainty	.95
Consumer Price Index for All Urban Consumers: Apparel	.88
New One Family Houses Sold: United States	.86
Industrial Production: Fuels	.82
Number Unemployed for 5-14 Weeks	.77
Continued Claims (Insured Unemployment)	.73
ISM Manufacturing: Supplier Deliveries Index	.72
Securities in Bank Credit, All Commercial Banks	.71
Industry Returns: Agricultural	.70
Money Zero Maturity: Money Stock	.69
Total Consumer Loans and Leases Owned and Securitized by Finance Companies	.66
Industrial Production: Residential Utilities	.62
Employment Cost Index: Benefits: Private Industry Workers	.61
Reserves of Depository Institutions, Nonborrowed	.58
Number Unemployed for Less than 5 Weeks	.53
Total Borrowings of Depository Institutions from the Federal Reserve	.53
Gross Saving	.53
Economic Policy Uncertainty: Government Spending	.50
M1 Money Stock	.46
Light Weight Vehicle Sales: Autos & Light Trucks	.44
Industry Portfolio Returns: Other	.43
Economic Policy Uncertainty: Sovereign Debt Currency Crises	.40
Fama-French Factor: Momentum	.38
All Employees, Government	.36
Civilian Employment Level	.35
Change in Private Inventories	.35
Industry Portfolio Returns: Drugs	.34
Nonperforming Commercial Loans	.34
Industry Portfolio Returns: Smoke	.33
CBOE NASDAQ 100 Volatility Index	.33
Consumer Sentiment Index	.30
Economic Policy Uncertainty: Trade policy	.29
Consumer Price Index for All Urban Consumers: Medical Care	.29
Industrial Production: Nondurable Consumer Goods	.28
National Income	.28
Number Unemployed for 15-26 Weeks	.27
Civilian Labor Force Level	.26
University of Michigan: Inflation Expectation	.26
Corporate Profits after Tax with IVA and CCAdj: Net Dividends	.26
Industrial Production: Materials	.25
Excess Reserves of Depository Institutions	.25

learners (linear or nonlinear) and the two other ensembles (bagged MARS and model-averaged NNs). Nonetheless, even if they are less competitive than rule-based ensembles, bagged MARS and model-averaged neural networks outperform individual learners in almost all model specifications.

This can be explained as follows. First, the prediction errors of ensemble methods generally have a lower variance than those of individual learners. This effect is actually a consequence of the aggregation of several base learners; it is particularly visible when comparing the performance of MARS against that of its bagged version. Second, rule-based ensembles are better suited to capture subgroups of data with similar properties and to build a separate model for each group. Recovery rates are particularly prone to behave

in this manner, as explained in Yao, Crook, and Andreeva (2015), Nazemi, Heidenreich, and Fabozzi (2018) and Bellotti, Brigo, Gambetti, and Vrins (2021). Third, rule-based methods are better suited to reproduce predictor-response relationships that are defined on a closed interval, as is generally the case for recovery rates.

It appears that the only individual learner that can compete with rule-based ensembles is the SVR. However, this method yields good performance only when trained using the full set of predictors or when systematic predictors are selected using Lasso without stability selection (which corresponds to the framework adopted in Nazemi and Fabozzi (2018)).

3.3.4 Meta-learning: within and across predictor sets

Let us now analyze the performance of linear meta-learning algorithms *within* each specification of the predictors set (highlighted in Figure 3.4 and Table 3.12). We notice a sharp drop in the average RMSE metrics for almost all specifications compared to the other models. In fact, the architecture of these algorithms allows the creation of more flexible functional forms thanks to the combination of various first-level learners. The different strengths of first-level learners are merged together to yield more accurate forecasts. The performance of linear meta-learning algorithms can be compared to those of the best rule-based ensembles. Indeed, Table 3.12 shows that OPT+, COS-E, and COS-IL always join the superior set of models with at least 10% probability outperforming the equally weighted forecast and the Hill Climbing algorithm, which, nevertheless, remain competitive, especially in specification 3,4 and 5.

The variation of the aggregate errors within linear meta-learners is much lower than within the class of individual or ensemble models, and the same holds true for the maximum average loss. Therefore, model risk is sensibly reduced when relying on meta-learners than on both individual learners and ensembles. This feature can be observed from Figure 3.4, and holds across all model specifications. We also find that nonlinear meta-learners (NNs) should generally be avoided. They usually display larger RMSE values than those of linear meta-learners across all specifications of the predictor set. Moreover, their partial advantage over ensemble methods and first-level learners in some specifications is neutralized by considerably higher errors in others (i.e., specifications 1, 8, and 9). Hence, in terms of model risk, ensembles, and linear meta-learners should be preferred.[14]

Nevertheless, in practice, we do not know ex-ante which among the nine specifications of the predictors set will offer the best performance. While we have previously employed meta-learning to learn the predictive set-up within each specification, we now use the same tools to combine all models *across* all specifications. In other words, we now jointly consider 180 base learners (20 models times 9 predictors specifications). Table 3.13 displays the RMSE, mean absolute error (MAE), and R^2 of the best 20 predictive strategies in this exercise. Boosted trees (specification 2, 4, 6, 7), Quantile random forest (specification 3), and Random forest (specification 3) are the only base learners that join the superior set of models. However, their performance clearly depends on the specification of the predictors set. Among meta-learners, while Opt suffers from the increased dimensionality of the problem and performs poorly, meta-learners with nonnegative combining weights (Opt+, COS-E, and COS-IL), occupy three of the top

[14]We find this conclusion to be robust to different specifications of the nonlinear meta-learner's architecture (i.e., the number of hidden units in the artificial neural networks). The results are available upon request.

Table 3.12: Out-of-sample root mean squared error (RMSE) metrics across model specifications. We indicate in **bold** the RMSE of the models that join the superior set of models with 10% α.

	1	2	3	4	5	6	7	8	9	Mean
Linear Models	.291	.204	.211	.211	.211	.210	.212	.213	.236	.222
Lin. regression	.774	.206	.205	.206	.206	.206	.206	.209	.233	.272
Lin. reg. back. sel.	.221	.205	.205	.207	.207	.206	.209	.214	.235	.212
Lasso 1	.203	.203	.206	.207	.207	.207	.208	.209	.233	.209
Lasso 2	.218	.203	.221	.221	.220	.218	.220	.221	.241	.220
Ridge 1	.202	.203	.208	.207	.207	.207	.208	.209	.233	.209
Ridge 2	.222	.211	.226	.220	.221	.220	.222	.223	.242	.223
Elastic net	.201	.202	.206	.207	.208	.207	.209	.209	.233	.209
Nonlinear Models	.211	.203	.207	.206	.211	.208	.207	.213	.237	.211
MARS	.207	.199	.214	.197	.202	.198	.194	.216	.237	.207
Gaussian processes	.211	.204	.205	.207	.207	.208	.207	.211	.234	.210
RVM	.236	.212	.200	.207	.223	.214	.216	.212	.230	.217
SVM	.191	.197	.212	.212	.212	.212	.211	.213	.245	.212
Rule-based Models	.232	.225	.233	.216	.217	.213	.221	.216	.240	.224
Regression tree	.230	.223	.244	.218	.215	.215	.222	.211	.256	.226
Conditional inference tree	.234	.226	.222	.214	.219	.211	.220	.221	**.225**	.221
Nonlinear Ensembles	.195	.197	.197	.196	.200	.200	.206	.206	.235	.203
Neural networks	.202	.201	.202	.199	.204	.207	.209	.208	.238	.208
Bagged MARS	.189	.194	.191	.193	.195	.193	.203	.204	.233	.199
Rule-based Ensembles	.189	.191	.186	.186	.187	.185	.183	.185	.229	.191
Cubist	**.184**	**.182**	**.184**	**.187**	.194	.186	.188	**.183**	**.229**	.191
Boosted trees s.g.b.	.194	.198	.195	.189	.187	.188	.189	.187	.238	.196
Boosted trees	**.182**	**.191**	**.174**	**.177**	**.185**	**.179**	**.177**	**.182**	**.223**	.185
Quantile random forests	.193	.195	.192	.192	.185	.187	**.180**	.187	.236	.194
Random forests	.190	**.187**	**.187**	**.185**	.185	.187	.182	.184	**.220**	.190
Linear Meta-Learning	.187	.191	.180	.185	.179	.184	.179	.179	.228	.188
Opt	**.188**	**.193**	**.181**	**.183**	**.180**	**.182**	**.177**	**.181**	**.227**	.188
Opt+	**.187**	**.191**	**.180**	**.186**	**.179**	**.184**	**.180**	**.178**	.228	.188
COS-E	**.187**	**.191**	**.180**	**.186**	**.179**	**.184**	**.180**	**.178**	**.228**	.188
COS-IL	**.187**	**.191**	**.180**	**.186**	**.179**	**.184**	**.180**	**.178**	**.228**	.188
Equally weighted for.	**.186**	**.187**	.188	**.187**	.188	.188	.187	.188	**.224**	.192
Hill Climbing	.193	.195	**.180**	.189	**.178**	.188	.189	.187	**.225**	.192
NonLinear Meta-Learning	.224	.200	.194	.194	.192	.191	.195	.213	.247	.205
NN - 1	.241	.198	.193	.187	.187	**.180**	.189	.234	.241	.205
NN - 2	.208	.203	.195	.201	.196	.201	.201	.192	.253	.206

four places and join the superior set of models, hence mitigating the uncertainty related to the choice of both the predictors set and the modeling framework, i.e., individual vs ensembles, linear vs nonlinear vs rule-based methods.

The results do not significantly differ when analyzing R^2 instead of RMSE; this is not surprising given the close relationship between the two metrics. The ranking remains coherent with previous results, especially among the top-performing methods: linear meta-learning techniques are still at the top, beaten only by Boosted trees in specification 7, i.e., when inflation uncertainty and Federal/State/local expenditures uncertainty, are considered together with industry default rate, commercial and industrial loans delinquency rates, industrial production, market index returns, PMI and individual characteristics.

Table 3.13: Ranking of top 20 predictive frameworks according to out-of-sample root mean squared error (RMSE) metrics across all specifications of the predictors set. The symbols $*$, $**$, and $***$ marks 1%, 5% and 10% significance levels of the model confidence test with L2 loss. For the sake of comparison, we also display the results corresponding to the equally weighted forecast and Hill Climbing algorithm despite ranking respectively 24th and 37th.

Model	Specification	RMSE	MAE	R^2	MCS
Boosted trees	7	.1735	.1076	.6298	***
Opt+	.	.1752	.1041	.6175	***
COS-E	.	.1752	.1041	.6175	***
COS-IL	.	.1752	.1041	.6175	***
Boosted trees	3	.1765	.1123	.6142	***
Boosted trees	6	.1771	.1140	.6115	***
Boosted trees	4	.1789	.1118	.6048	***
Quantile random forests	3	.1796	.0963	.6080	***
Boosted trees	2	.1819	.1117	.5920	***
Random forests	3	.1820	.1140	.5927	***
Boosted trees	9	.1823	.1140	.5886	
Cubist	8	.1824	.1053	.5878	
Cubist	2	.1831	.1053	.5858	
Cubist	7	.1836	.1076	.5841	
Cubist	9	.1836	.1029	.5843	
Random forests	2	.1842	.1136	.5818	
Opt	.	.1843	.1078	.5852	
Boosted trees	5	.1845	.1205	.5781	
Random forests	6	.1849	.1164	.5869	
Quantile random forests	5	.1853	.0964	.5887	
Equally weighted for.	.	.1862	.1289	.5962	
Hill Climbing	.	.1890	.1213	.5622	

Nevertheless, this difference in performance is not statistically significant: COS-E, COS-IL, and Opt+ join the superior set of models with a 10% confidence level. Conversely, equally weighted forecast and the Hill Climbing algorithm perform poorly and are both excluded from the superior set of models. The main difference that emerges when considering the MAE is that quantile random forests (QRF) outperform their competitors, on average. This should not come as a surprise: QRF are trained to minimize the MAE to mitigate the impact of outliers on estimation error, while the loss functions of other predictive strategies are instead related to the RMSE. Overall, COS-E, COS-IL, and Opt+ still provide very competitive MAE, while, at the same time, mitigating the uncertainty related to the choice of both the predictors set and the modeling framework.

3.4 Discussion and practical considerations

The results of the previous subsections convey important practical considerations that we now summarize.

First, financial institutions should accelerate the implementation of LGD internal models instead of maintaining the use of historical averages. We show in Subsection 3.3.1 that the predictive model does not need to be complicated to be effective. All model specifications outperform the historical average approach, which, however, is the method underlying the standard and foundation IRB frameworks. Unfortunately, recent regulatory guidelines (BCBS, 2017) are now pointing in the opposite direction, favoring fixed recovery rate approaches. Our results suggest that this is perhaps not the best route to follow, as LGD internal models can produce more reliable risk figures.

Second, in Subsection 3.3.2, we confirm the necessity of including economic factors that can capture systematic fluctuations in recovery rates. This result is in line with the current regulatory prescriptions. Models relying on well-identified economic determinants provide remarkable results, while *big data* and variable selection procedures do not necessarily translate into better performance because results strongly depend on the considered predictive algorithms. We found that financial uncertainty and text-based economic uncertainty measures are relevant predictors of recovery rate ex-ante. This extends the finding of Gambetti, Gauthier, and Vrins (2019), where financial uncertainty and the Economic Policy Uncertainty Index were found to explain ex-post recovery rates. Similarly, we also confirm the finding of Nazemi and Fabozzi (2018): inflation measures, inflation expectation, industrial production, corporate profits after tax, indicators of the housing market, and stock market volatility are not only systemic determinants of recovery rates but also important predictors. This is also confirmed when using Lasso with stability selection (Table 3.11): the probability of including Financial Uncertainty and text-based economic policy uncertainty measures is particularly high. Specifically, Financial Uncertainty, Economic Policy Uncertainty: Government Spending, Economic Policy Uncertainty: Sovereign Debt Currency Crises, and Economic Policy Uncertainty: Trade policy have 95%, 50%, 40%, and 29% probability of being selected, respectively. Including such predictors in more parsimonious and theoretically justified models makes the predictive setting easier to implement, interpret, and back-test, making it more prone to be validated by regulators.

Third, in a context where increasing amounts of data become publicly available, the common practice of using standard linear regression to forecast recovery rates should be abandoned. In Subsection 3.3.1, we confirm that this is incompatible with the possibility of exploiting the information contained in a large data set of candidate predictors and this embeds high model risk. This result is in line with the recent findings of Dong, Yan, Rapach, and Guofu (2020) in a similar financial context.

Fourth, we show in Subsection 3.3.3 that ensemble methods and meta-learning techniques should be fostered in the industry. The choice of the A-IRB model should be directed towards ensemble methods with respect to individual learners, as they always yield better performance. However, an incorrect choice among ensembles may still lead to worse results than an incorrect choice among meta-learning techniques. Hence, meta-learning considerably reduces model risk compared to ensembles; it should be considered by the regulator for a new set of A-IRB guidelines. This is particularly true for linear meta-

learning techniques that also have the advantage of providing interpretable insights about the contribution of each individual model. As they show, the benefits of diversification can already be appreciated by combining a limited number of high-performing algorithms.

3.5 Conclusion

In this chapter, we explore for the first time the applicability and the potential of *meta-learning* in order to forecast bond recovery rates. Meta-learning consists of combining the predictions arising from several first-level algorithms into a single aggregated forecast. More specifically, our purpose is to test the performance of a wide set of machine learning methods for predicting recovery rates on defaulted corporate bonds using a data set of predictors of unprecedented size and see whether combining them might display superior performance. We consider as first-level algorithms both individual and ensemble methods belonging to the classes of linear, nonlinear, and rule-based models.

We contribute to the literature on bond recovery rate prediction in three ways. Our first contribution deals with the set of predictors to use. We find that including a limited number of well-identified recovery rate determinants and economic uncertainty yields better predictive performance than using the full set of macroeconomic predictors, even when combined with variable selection techniques such as Lasso. This finding prevails in all considered algorithms and contrasts with the current Big Data trend. The use of a limited number of economically grounded predictors offers the additional advantages of making the model easier to implement, to back-test and therefore, when applicable, to be validated by regulatory instances because of mitigating the risk of over-fitting and data-mining. Interestingly, we confirm the central importance of uncertainty measures, which are always retained by the considered variable selection methods.

The second and third contributions relate to the predictive method to use. Random forests, quantile random forests, boosted trees, and cubist display the most promising results but their performance is unstable across the considered specifications of the predictors set: it seems difficult to identify ex-ante the right pair of model and predictors set to use. By contrast, we empirically show that meta-learning can beat this challenge. Indeed, meta-learning can improve recovery rate predictions compared to traditional individual learning machines while, at the same time, considerably reducing model risk. This evidence is preserved both when looking at the predictive performance within a chosen predictor set and when considering jointly predictions across all specifications of the predictor set.

In all specifications, the historical average approach performs significantly worse. Yet, this is the method underlying the standardized approach and foundation F-IRB framework used to compute regulatory capital reserves, which is the primary figure used worldwide to monitor the creditworthiness of banks. Therefore, our findings suggest that regulators and policymakers promoting the use of more reliable risk figures should foster the implementation of LGD internal models that use meta-learning techniques, while the use of traditional linear regressions or historical averages should be challenged. Eventually, our findings are of practical interest to design new tools for internal economic capital calculations as well as for stress testing purposes.

APPENDIX

A3.1 Details on Nonlinear and rule-based methods

Following Bellotti, Brigo, Gambetti, and Vrins (2021), we give more details on the nonlinear and rule-based methods we have considered in this study. We refer to Hastie, Tibshirani, and Friedman (2009) for a textbook treatment of the predictive algorithms.

Multivariate adaptive regression splines (MARS)

The multivariate adaptive regression splines method of Friedman (1991) features piece-wise transformations of the original predictive variables. Specifically, given a set of cut-points t, each predictor X_j is transformed into a set of *reflected pairs* obtained by:

$$(X_j - t)_+ = \begin{cases} X_j - t, & \text{if } X_j > t \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad (t - X_j)_+ = \begin{cases} t - X_j, & \text{if } X_j > t \\ 0, & \text{otherwise} \end{cases} \tag{A3.1}$$

where $j = (1, 2, ..., p)$ and $t \in \{x_{1j}, x_{2j}, ..., x_{Nj}\}$. A linear regression model is then estimated following a greedy procedure on the reflected pairs that are selected for each predictor. The number of selected pairs for each predictor defines the *degree* of the MARS algorithm. A backward *pruning* procedure handles over-fitting by dropping the features that are associated with the smallest error rate when excluded. In this study, we consider MARS of degree one and the number of terms to be retained in the final model is estimated via 10-fold cross-validation.

K-nearest neighbors

K-nearest neighbors is a non-parametric method that forecasts the target variable by using the average of the K training observations that are closest in the covariate space. *Closeness* between observations is defined by the Euclidean distance. In this chapter, we tune the size of the neighborhood K via 10-fold cross-validation.

Model averaged neural networks

Model averaged neural networks (Ripley, 1996) is an ensemble of feed-forward neural networks where the base learners use different initial values for the optimization procedure to estimate the parameters. Beside reducing the model risk via averaging the multiple predictions of the base learners, the random initialization of the parameters mitigates the risks of converging in local optima when dealing with back-propagation algorithm. Individual neural networks can also include a *weight decay* that penalizes large coefficients further mitigating the risk of over-fitting. We employ an ensemble of 5 neural networks and we tune the number of hidden units as well as the weight decay via 10-fold cross-validation.

Support vector regression, relevance vector machine and Gaussian processes

Support vector regression (SVR) (Vapnik, 1995) is a kernel-based model estimated from the following minimization problem:

$$\underset{\beta_0,\,\beta}{\arg\min}\; C\sum_{i=1}^{N} L_\epsilon(y_i - f(x_i)) + \sum_{j=1}^{p}\beta_j^2\;, \tag{A3.2}$$

where L_ϵ is an ϵ-insensitive loss function and C is the penalty assigned to residuals greater or equal to ϵ. The solution to this problem ((A3.2)) can be expressed in terms of a set of weights w_i, and a positive definite kernel function $K(\cdot)$ that depends on the training set data points:

$$f(x) = w_0 + \sum_{i=1}^{N} w_i K(x,\, x_i)\;. \tag{A3.3}$$

Training observation associated to non-zero weights are called *support vectors* and they are used to estimate the model. We consider the radial basis kernel $K(x,\, x') = \exp(-\sigma\|x - x'\|)$ and we estimate the penalty C via 10-fold cross-validation while the scaling parameter σ following Caputo, Sim, Furesjo, and Smola (2002).

Relevance vector machine (RVM) (Tipping, 2001) is a kernel method similar to SVR, but the weights are now computed using a Bayesian framework. This allows for a probabilistic interpretation of the model predictions and it offers the additional advantage of producing sparser models than standard SVR.

Williams and Rasmussen (1996) introduce Gaussian processes, a non-sparse non-parametric generalization of the RVM framework. Gaussian processes impose a prior distribution directly on the function values and consider Gaussian distribution with zero mean and covariance matrix equal to the kernel matrix $K_{ij} = K(x_i, x_j)$. In this work, we implement Gaussian processes using a radial basis kernel.

Regression trees

Regression trees (Breiman, Friedman, Olshen, and Stone, 1984) partition the predictors' space into a set of non-overlapping regions and fit a model in each of them. In their simplest version, predictions are formed using the average of the target variables associated with each region. The algorithm employs a top-down partitioning to estimate the regions, starting with the full dataset and sequentially splitting it into groups according to the predictors and cut-points that achieve the largest decrease in the residual sum of square errors. The splitting procedure stops when a certain criterion is met, e.g., the number of observations in the terminal nodes. To limit over-fitting, the tree is then pruned back by using *cost and-or complexity* criterion and the amount of regularization can be tuned via cross-validation.

Conditional inference trees

The conditional inference trees algorithm (Hothorn, Hornik, and Zeileis, 2006) was proposed to overcome the potential selection bias of regression trees. In fact, in the standard regression trees algorithm there is a greater chance of selecting the predictors featuring a

higher number of candidate cut points during the tree growing step. For each predictor, conditional inference trees use statistical hypothesis testing to evaluate the difference between the means of the two samples created by a candidate split. Multiple comparison corrections reduce the selection bias for highly granular predictors (Westfall and Young, 1993). In this work, we estimate the p-value threshold determining whether considering a new split and the tree maximum depth via 10-fold cross-validation.

Bagged trees and random forests

Bagged trees (Breiman, 1996b) is an ensemble method resulting from the aggregation of the predictions of multiple regression trees estimated on bootstrapped samples of the training set. In this work, we tune the number of trees composing the ensemble by 10-fold cross-validation.

Random forests (Breiman, 2001) is an *evolution* of the bagged trees algorithm that was motivated by overcoming the problem of high correlation between individual trees. In fact, for bagged trees all predictors are considered at each split during the tree-growing process, which results in trees grown in very similar structures. The random forests algorithm instead considers a subset of randomly selected predictors is considered at each split. This reduces the correlation between the trees and also the variance of the ensemble prediction. We tune the number of randomly selected predictors via 10-fold cross-validation.

Boosted trees

Boosted trees is an ensemble algorithm where the individual trees are *sequentially* fitted by multiple weak learners and, to avoid over-fitting, only a percentage of each fitted value (called *learning rate*) is subtracted from the residual from the previous learner. The number of boosting iterations, i.e., the number of trees, the learning rate and the individual trees' depth are the main hyper-parameters for this model. Stochastic gradient boosting is a version of the algorithm that includes a random sampling scheme of the training data at each iteration to improve computational efficiency and further mitigating the risk of over-fitting. In this work, we tune the hyper-parameters via 10-fold cross-validation.

Cubist

Cubist (Quinlan, 1993) combines the M5 model tree model of Quinlan (1992) with some features from boosting and K-nearest neighbors algorithms. Cubist has a tree structure where each node contains a linear regression model whose covariates are the same variables that satisfy the rule defining a specific node. After the model is estimated, the predictions in each node are recursively *smoothed* by using the fitted values from the respective parent node. Smoothing consists of linearly combining the two models with the one with the smallest RMSE having the largest weight. To limit over-fitting, the rules can be pruned using the adjusted error rate criterion as in M5. Cubist can also adjust the forecast using a weighted average of sample neighbors. Committees, i.e., ensembles, can also be created using multiple model trees in a boosting-like framework. In this work, we tune the number of neighbors and the number of committees via 10-fold cross-validation.

A3.2 A closer look at uncertainty measures

Following Gambetti, Gauthier, and Vrins (2019), we provide more details on the financial and economic uncertainty measures employed in this study.

We consider two measures of financial uncertainty: a) stock market volatility and b) the financial uncertainty index of Jurado, Ludvigson, and Ng (2015a). On the one hand, the stock market volatility-based measure of financial uncertainty included in our data set is represented by the stock market implied volatility index VIX, retrieved from the Chicago Board Options Exchange database. It has monthly frequency and it is also included in the FRED monthly database. On the other hand, we select the one-month horizon financial uncertainty indicator of Jurado, Ludvigson, and Ng (2015a) and Ludvigson, Ma, and Ng (2015) that is instead based on forecast error volatility. This latter method identifies uncertainty with unpredictability, i.e, the more or less the macroeconomic series have become predictable, the less or more uncertainty agents face. Specifically, Jurado, Ludvigson, and Ng (2015a) define h-period ahead uncertainty in the variable Y to be the conditional volatility of the purely unforecastable (at horizon h) component of the future value of the series. Data was retrieved from the authors' website. Uncertainty measures can also be based on text-analysis. This is the case of the economic policy uncertainty index and its sub-components from Baker, Bloom, and Davis (2016). The series, available monthly, are retrieved from `https://www.policyuncertainty.com/` and include a range of sub-indexes based solely on news data. Specifically, these are extracted from the Access World News database of over 2,000 US newspapers and measure coverage frequency. For example, each sub-index measures the coverage frequency of economic, uncertainty, and policy terms as well as a set of categorical policy terms, e.g., monetary policy, tax government spending or trade policy (see the news-based measures in Table 3.9 for the complete list) in the pool of articles under analysis. For instance, articles that fulfill the requirements to be coded as economic policy uncertainty and also contain the term 'federal reserve' would be included in the monetary policy uncertainty sub-index. For more details, please refer to Appendix B of Baker, Bloom, and Davis (2016). We also consider survey-based measures that are also available at `https://www.policyuncertainty.com/`. The first, drawing on reports by the Congressional Budget Office, reflects the number of federal tax code provisions set to expire over the next 10 years. The second, relying on the Federal Reserve Bank of Philadelphia's Survey of Professional Forecasters, measures the dispersion between individual forecasters' predictions about future levels of the Consumer Price Index, Federal Expenditures, and State and Local Expenditures among economic forecasters as a proxy of uncertainty about policy-related macroeconomic variables.

A3.3 FRED data

The list of the variables considered in this study is provided below. We indicate with Δ the differencing order and with *log* the natural logarithm operator. The column tcode denotes the following data transformation for a series x: (1) no transformation; (2) Δx_t; (3) $\Delta 2 x_t$; (4) $log(x_t)$; (5) $\Delta log(x_t)$; (6) $\Delta 2 log(x_t)$; (7) $\Delta(\frac{x_t}{x_{t-1}} - 1)$. For more details on the calculation of the variables denoted with (*), please reefre to the data appendix of Jurado, Ludvigson, and Ng (2015a) .

Table A3.1: List of Macroeconomic Predictors, their code and abbreviation in FRED and its considered tranformation.

Variable	Description	tcode
COUP_RATE	Coupon Rate	1
BACK_F	Presence of additional backing guarantees different from the issuer's assets	1
DEF_DEBT_SENR.Senior.Subordinated	Seniority Status	1
DEF_DEBT_SENR.Senior.Unsecured	Seniority Status	1
DEF_DEBT_SENR.Subordinated	Seniority Status	1
MOODYS_11_CODE.Capital.Industries	Industry Code	1
MOODYS_11_CODE.Consumer.Industries	Industry Code	1
MOODYS_11_CODE.Energy...Environment	Industry Code	1
MOODYS_11_CODE.FIRE	Industry Code	1
MOODYS_11_CODE.Media...Publishing	Industry Code	1
MOODYS_11_CODE.Retail...Distribution	Industry Code	1
MOODYS_11_CODE.Technology	Industry Code	1
MOODYS_11_CODE.Transportation	Industry Code	1
MOODYS_11_CODE.Utilities	Industry Code	1
DEF_TYP_CD.Missed.interest.payment	Deafult Type	1
DEF_TYP_CD.Missed.principal.and.interest.payments	Deafult Type	1
DEF_TYP_CD.Missed.principal.payment	Deafult Type	1
DEF_TYP_CD.Others	Deafult Type	1
DEF_TYP_CD.Prepackaged.Chapter.11	Deafult Type	1
DEF_TYP_CD.Receivership	Deafult Type	1
DEF_TYP_CD.Suspension.of.payments	Deafult Type	1
FinUnc_h.1	Financial Uncertainty	1
Baseline_overall_index	Economic Policy Uncertainty index	1
News_Based_Policy_Uncert_Index	Newspaper Based Policy Uncertainty	1
FedStateLocal_Ex_disagreement	Federal Tax Code Uncertainty	1
CPI_disagreement	CPI Survey Disagreement	1
X1..Economic.Policy.Uncertainty	1. Economic Policy Uncertainty	1
X2..Monetary.policy	2. Monetary policy	1
Fiscal.Policy..Taxes.OR.Spending.	Fiscal Policy (Taxes OR Spending)	1
X4..Government.spending	4. Government spending	1
X5..Health.care	5. Health care	1
X6..National.security	6. National security	1
X7..Entitlement.programs	7. Entitlement programs	1
X8..Regulation	8. Regulation	1
Financial.Regulation	Financial Regulation	1
X9..Trade.policy	9. Trade policy	1
X10..Sovereign.debt..currency.crises	10. Sovereign debt, currency crises	1
RPI	Real Personal Income	5
W875RX1	Real personal income excluding R current transfer receipts	5
DPCERA3M086SBEA	Real personal consumption expenditures	5
CMRMTSPLx	Real Manu. and Trade Industries Sales	5
RETAILx	Retail and Food Services Sales	5
IPFINAL	IP: Final Products (Market Group)	5
IPCONGD	IP: Consumer Goods	5
IPDCONGD	IP: Durable Consumer Goods	5
IPNCONGD	IP: Nondurable Consumer Good	5
IPBUSEQ	IP: Business Equipment	5
IPMAT	IP: Materials	5
IPDMAT	IP: Durable Materials	5
IPNMAT	IP: Nondurable Materials	5
IPB51222S	IP: Residential Utilities	5
IPFUELS	IP: Fuels	5
CUMFNS	Capacity Utilization: Manufacturing	2

Table A3.2: List of Macroeconomic Predictors, their code and abbreviation in FRED and its considered tranformation.

Variable	Description	tcode
HWI	Help-Wanted Index for U.S.	2
HWIURATIO	Ratio of Help Wanted/No. Unemployed	2
CLF16OV	Civilian Labor Force	5
CE16OV	Civilian Employment	5
UNRATE	Civilian Unemployment Rate	2
UEMPMEAN	Average Duration of Unemployment (Weeks)	2
UEMPLT5	Civilians Unemployed - Less Than 5 Weeks	5
UEMP5TO14	Civilians Unemployed for 5-14 Weeks	5
UEMP15OV	Civilians Unemployed - 15 Weeks & Over	5
UEMP15T26	Civilians Unemployed for 15-26 Weeks	5
UEMP27OV	Civilians Unemployed for 27 Weeks and Over	5
CLAIMSx	Initial Claims	5
PAYEMS	All Employees: Total nonfarm	5
CES1021000001	All Employees: Mining and Logging: Mining	5
USCONS	All Employees: Construction	5
DMANEMP	All Employees: Manufacturing	5
NDMANEMP	All Employees: Nondurable goods	5
USWTRADE	All Employees: Wholesale Trade	5
USTRADE	All Employees: Retail Trade	5
USFIRE	All Employees: Financial Activities	5
USGOVT	All Employees: Government	5
CES0600000007	Avg Weekly Hours : Goods-Producing	1
AWOTMAN	Avg Weekly Overtime Hours : Manufacturing	2

Table A3.3: List of Macroeconomic Predictors, their code and abbreviation in FRED and its considered tranformation.

Variable	Description	tcode
M1SL	M1 Money Stock	6
M2REAL	Real M2 Money Stock	5
AMBSL	St. Louis Adjusted Monetary Base	6
TOTRESNS	Total Reserves of Depository Institutions	6
NONBORRES	Reserves Of Depository Institutions	7
BUSLOANS	Commercial and Industrial Loans	6
REALLN	Real Estate Loans at All Commercial Banks	6
NONREVSL	Total Nonrevolving Credit	6
CONSPI	Nonrevolving consumer credit to Personal Income	2
S.P..indust	S&P's Common Stock Price Index: Industrials	5
S.P.div.yield	S&P's Composite Common Stock: Dividend Yield	5
S.P.PE.ratio	S&P's Composite Common Stock: Price-Earnings Ratio	5
FEDFUNDS	Effective Federal Funds Rate	2
CP3Mx	3-Month AA Financial Commercial Paper Rate	2
TB3MS	3-Month Treasury Bill	2
TB6MS	6-Month Treasury Bill	2
GS1	1-Year Treasury Rate	2
GS5	5-Year Treasury Rate	2
AAA	Moody's Seasoned Aaa Corporate Bond Yield	2
BAA	Moody's Seasoned Baa Corporate Bond Yield	2
COMPAPFFx	3-Month Commercial Paper Minus FEDFUNDS	1
TB3SMFFM	3-Month Treasury C Minus FEDFUNDS	1
TB6SMFFM	6-Month Treasury C Minus FEDFUNDS	1
T1YFFM	1-Year Treasury C Minus FEDFUNDS	1
T10YFFM	10-Year Treasury C Minus FEDFUNDS	1
AAAFFM	Moody's Aaa Corporate Bond Minus FEDFUNDS	1
BAAFFM	Moody's Baa Corporate Bond Minus FEDFUNDS	1
TWEXMMTH	Trade Weighted U.S. Dollar Index: Major Currencies, Goods	5
EXSZUSx	Switzerland / U.S. Foreign Exchange Rate	5
EXJPUSx	Japan / U.S. Foreign Exchange Rate	5
EXUSUKx	U.S. / U.K. Foreign Exchange Rate	5
EXCAUSx	Canada / U.S. Foreign Exchange Rate	5
WPSFD49207	PPI: Finished Goods	6

Table A3.4: List of Macroeconomic Predictors, their code and abbreviation in FRED and its considered tranformation.

Variable	Description	tcode
WPSFD49207	PPI: Finished Goods	6
WPSID62	PPI: Crude Materials	6
OILPRICEx	Crude Oil, spliced WTI and Cushing	6
PPICMM	PPI: Metals and metal products	6
CPIAPPSL	CPI : Apparel	6
CPIMEDSL	CPI : Medical Care	6
CUSR0000SAD	CPI : Durables	6
CUSR0000SAS	CPI : Services	6
DDURRG3M086SBEA	Personal Cons. Exp: Durable goods	6
DNDGRG3M086SBEA	Personal Cons. Exp: Nondurable goods	6
DSERRG3M086SBEA	Personal Cons. Exp: Services	6
CES0600000008	Avg Hourly Earnings : Goods-Producing	6
CES2000000008	Avg Hourly Earnings : Construction	6
CES3000000008	Avg Hourly Earnings : Manufacturing	6
UMCSENTx	Consumer Sentiment Index	2
MZMSL	MZM Money Stock	6
DTCOLNVHFNM	Consumer Motor Vehicle Loans Outstanding	6
DTCTHFNM	Total Consumer Loans and Leases Outstanding	6
INVEST	Securities in Bank Credit at All Commercial Banks	6
DelRate.ConsumerLoans	Delinquency rates on Consumer Loans	1
DelRate.CreditCardLoans	Delinquency rates on Credit Card Loans	1
DelRate.CommIndLoans	Delinquency rates on Commercial and Industrial Loans	1
AMR.Def..Rate	American Default Rate	1

Table A3.5: List of Macroeconomic Predictors, their code and abbreviation in FRED and its considered tranformation.

Variable	Description	tcode
D_log.DIV.	CRSP - Dividends *	5
D_Preinvested	CRSP - Price under reinvestment *	5
d.p	CRSP - Dividend to Price *	5
R15.R11	FF factor (Small, High) minus (Small, Low) sorted on (size, book-to-market)	1
CP.factor	FF factor - Cash Profitability	1
SMB	FF factor - Small - Big	1
UMD	FF factor - Momentum	1
Agric	Portfolio Return	1
Food	Portfolio Return	1
Beer & Liquor	Portfolio Return	1
Smoke	Portfolio Return	1
Toys - Recreation	Portfolio Return	1
Fun - Entertaiment	Portfolio Return	1
Books - Printing and Publishing	Portfolio Return	1
Hshld - Consumer Goods	Portfolio Return	1
Clths - Apparel	Portfolio Return	1
MedEq - Medical Equipment	Portfolio Return	1
Drugs - Pharmaceutical Products	Portfolio Return	1
Chems - Chemicals	Portfolio Return	1
Rubbr - Rubber and Plastic Products	Portfolio Return	1
Txtls - Textiles	Portfolio Return	1
BldMt - Construction Materials	Portfolio Return	1
Construction	Portfolio Return	1
Steel	Portfolio Return	1
Machinery	Portfolio Return	1
Electrical Equipment	Portfolio Return	1
Autos - Automobiles and Trucks	Portfolio Return	1
Aero - Aircraft	Portfolio Return	1
Ships	Portfolio Return	1
Mines - Non-Metallic and Industrial Metal Mining	Portfolio Return	1
Coal	Portfolio Return	1
Oil	Portfolio Return	1
Util - Utilities	Portfolio Return	1
Telcm - Communication	Portfolio Return	1
PerSv - Personal Services	Portfolio Return	1
BusSv - Business Services	Portfolio Return	1
Hardw - Computers	Portfolio Return	1
Chips - Electronic Equipment	Portfolio Return	1
LabEq - Measuring and Control Equipment	Portfolio Return	1
Paper - Business Supplies	Portfolio Return	1
Boxes - Transportation	Portfolio Return	1
Trans - Transportation	Portfolio Return	1
Whlsl - Wholesale	Portfolio Return	1
Rtail - Retail	Portfolio Return	1
Meals - Restaurants, Hotels, Motels	Portfolio Return	1
Banks - Banking	Portfolio Return	1
Insur - Insurance	Portfolio Return	1
RlEst - Real Estate	Portfolio Return	1
Fin - Trading	Portfolio Return	1
Other	Portfolio Return	1

Table A3.6: List of Macroeconomic Predictors, their code and abbreviation in FRED and its considered tranformation.

Variable	Description	tcode
A032RC1A027NBEA	National Income	5
HOUSTNE	Housing Starts, Northeast	4
HOUSTW	Housing Starts, West	4
ACOGNO	New Orders for Consumer Goods	5
AMDMNOx	New Orders for Durable Goods	5
ANDENOx	New Orders for Nondefense Capital Goods	5
AMDMUOx	Unfilled Orders for Durable Goods	5
BUSINVx	Total Business Inventories	5
ISRATIOx	Total Business: Inventories to Sales Ratio	2

Chapter 4

Business cycle and realized losses in the consumer credit industry

with Walter Distaso and Frédéric Vrins

Abstract

We analyze the determinants of loss given default (LGD) of consumer credit. Previous studies conclude that only contract-specific variables such as exposure-at-default or age matter; in contrast with defaulted corporate bonds, macro variables like GDP or business cycle are irrelevant. Exploiting a much larger dataset including more than 6 million of Italian consumer debt from 2007 to 2019, considering more than 40 predictors and a variety of models covering standard regressions, machine learning techniques, and forecasts combination schemes, we consistently obtain opposite results. We find that regional social and welfare indicators and credit cycle variables contribute to enhancing forecasting performance by up to 15 percentage points in terms of R^2. The relationships between the expected LGD and the macro predictors unveiled by accumulated local effects plots confirm the intuition that lower real activity, increasing cost-of-debt to GDP ratio, and greater economic uncertainty are associated with a greater LGD for consumer credit, too. In addition, our analysis suggests that the most effective approach for predicting LGDs for both individual credits and portfolio of credits is to use forecasts combination schemes. Among these methods, forecasts combinations based on artificial neural networks stand out. Using model confidence set theory, we find that they consistently outperform the other considered models in terms of mean absolute error and root mean squared error, always joining the superior set of models. Our results are important for credit institutions, third-party debt collection firms, or regulatory bodies in charge of capital requirements policies.

Introduction

Consumer credit (also known as consumer debt) refers to the debt that individuals contract to purchase consumption goods. This includes usual credit lines (issued by credit institutions or by the sellers directly) as well as liabilities resulting from unscheduled deferred payments.

Consumer credit is a fast-growing industry worldwide. In September 2022, for instance, consumer debt rose to USD 4,700.821 bln in the US and to 715.106 bln EUR in the euro area. This represents a growth of 76.6% in the US and 11.74% in the euro area concerning

levels observed before the outburst of the great financial crisis in September 2008. In the Eurozone, it accounts for 10.8% of the outstanding credit for households.[1]

Consumer credit is primarily concerned with credit risk, which arises when consumers fail to repay their debt, resulting in losses for creditors. The degree of loss suffered by the creditors can be measured using two parameters, namely the exposure at default (EAD) and the recovery rate (RR) (or loss given default, LGD = 1 - RR). The EAD refers to the outstanding principal amount of the debt at the time of default, while the RR denotes the ratio of the amount that creditors will recover to the EAD, hence, takes value in $[0, 1]$. The EAD is typically known at the time of default, whereas the RR determines the magnitude of the losses incurred by creditors when closing the debt collection process.

Confronted with a non-performing credit (NPC), a debt holder has four options: take her loss, initiate a collection procedure with the customer, give a mandate to a specialized firm that will collect the debt in her name, or sell the NPC at a discount to a company who will then try to recover part of the EAD by chasing the customer. To evaluate the relevance of initiating a collection procedure, to determine the fees of a third-party debt collection firm, or to price the NPC to be sold or bought, every stakeholder needs to estimate what the RRs are likely to be.

Although a vast literature deals with the RRs of defaulted corporate bonds, little is known about the RRs of consumer credits. However, these two types of RRs are not quite comparable. For instance, the assets are very different: bonds are issued by large institutions, hence, are much more liquid than individual or SME loans. In addition, both the amount and the nature (public vs private) of available information about them are different. Finally, it is well-known that the empirical distribution of the RRs of defaulted corporate bonds is smooth and unimodal, while that of NPCs exhibits a pronounced bimodal pattern (see, e.g., (Gambetti, Gauthier, and Vrins, 2019; Nazemi, Baumann, and Fabozzi, 2022; Nazemi, Rezazadeh, Fabozzi, and Höchstötter, 2022)). These differences suggest that one can hardly extrapolate the findings related to defaulted corporate bonds to NPCs. As detailed in the next section, most of the studies on defaulted corporate bonds and consumer credit RRs highlight that contract-specific variables (such as the seniority or the sector for bonds, and the maturity or the principal for consumer credit) matter. However, the impact of macroeconomic and social (MS) indicators on RR is less clear and seems to depend on the type of debt at hand. For instance, it is known since the seminal paper of Altman, Brady, Resti, and Sironi (2005) that realized recovery rates of corporate bonds are negatively correlated with the default rate, showing that macroeconomic variables matter for these assets. Surprisingly enough, this would not apply to consumer credit, where macroeconomic variables were shown to be irrelevant. This contrasts with the idea that income and the residual value of consumers' personal assets depend on the environment in which the recovery process takes place. In such a case, if realized recovery rates display pro-cyclical dynamics, consumer debt owners will risk facing greater realized losses than expected.

In our view, four main reasons can be put forward to explain this paradox. First, it is very difficult to access a meaningful dataset: data is often scarce and limited to a specific sector of the consumer credit industry so the findings can be hardly generalized.

[1]Data on credit extended to individuals for household, family, and other personal expenditures, excluding loans secured by real estate for U.S. is available at `https://www.federalreserve.gov/releases/g19/about.htm` and for the euro area can be retrieved from the MFI balance sheets available from the ECB Statistical Data Warehouse `https://sdw.ecb.europa.eu/`.

Second, the samples under analysis are relatively short (in the time dimension) while the economic indicators are low-frequency time series that evolve across business cycles. The low variability of MS indicators in short samples may explain their marginal relevance for describing consumer debt RRs compared to more granular credit-specific features such as exposure at default or debtor's age. Third, researchers have focused on RRs at individual credit levels, despite defaulted consumer credits are often evaluated and managed at a portfolio scale. For instance, third-party collectors acquire the right to recover for pools of defaulted credits, earning the totality (if purchased) or a fee (if handled in the name of the debt owner) on the amount recovered. In both cases, they aim at maximizing the total recovered amount and, therefore, studying the determinants at the portfolio level better reflects the business model of this industry. Moreover, while MS indicators may have little predictive power at the individual level, they may play a bigger role at the pool level, where idiosyncratic risk is diversified. Fourth, recovery rate modeling has mostly focused on in-sample *vs* out-of-sample splits to analyze recovery rate determinants or discuss the best predictive strategies. Nevertheless, when the data is randomly assigned to in-sample and out-of-sample sets, realizations of RRs used for training the models have likely occurred *after* the defaults of credits used for evaluating the models. The underlying assumption is that the data-generating process is time-invariant which, if not met, can potentially lead to a look-ahead bias in the predictions, therefore invalidating the findings.

This chapter investigates the relevance of macroeconomic and social variables in explaining recovery rates and therefore realized losses in the broader consumer credit industry by relying on a real portfolio of NPCs owned by a debt collector.

We make the following contribution to the literature on consumer credit. First, we analyze a database counting more than 6 million NPCs from utilities, banking, financial and commercial sectors in Italy, spanning the years 2007 - 2020. To the best of our knowledge, this is the largest dataset considered so far in the academic literature. Second, we study the determinants of LGDs both at the individual NPC levels but also at portfolio levels. While existing studies focus on the former, the latter also matters when, for example, the problem at hand is to determine the fair price of a portfolio of NPCs to be transferred from one party to another. Third, exploiting the heterogeneous economic dynamics across regions and provinces of Italy[2], the data permits us to understand how the economy impacts LGDs in the consumer credit industry. In contrast with previous studies analyzing the determinants of NPC using smaller datasets focusing on specific types of NPC, we find that excluding MS indicators from the set of predictors significantly harms forecasting performance both at individual credit and portfolio levels. We believe that this effect comes from the lack of variability in macro indicators when considering smaller time frames. Our findings are robust to the choice of forecasting models; they remain valid when considering standard regressions, machine learning algorithms, and robust combinations of those. Fourth, when it comes to time-consistent prediction exercises, our conclusion becomes even more significant. Discarding the time dimension, including MS variables improves the R^2 for beta and linear regression models by 4 percentage points. In the inter-temporal framework, this number jumps to 15 percentage points. Accumulated

[2]For instance, in 2018, the GDP per capita ranged from EUR 23,879 in Sicily to EUR 35,968 in Lombardy, with the average for Italy being EUR 31,641. Similarly, during the first half of 2020, credit to firms increased by varying percentages in different regions, with a higher increase in the Center, North-West, and North-East regions and a lower increase in the Islands and South of Italy, according to (Bank of Italy, 2020).

local effect plots (ALE) confirm that our results are in line with intuition. For instance, a deterioration of the credit conditions at the national level is associated with higher LGDs for portfolios of NPCs.

The chapter is structured as follows. In Section 4.1, we review the salient results related to the determinant of RRs for various types of debts. Section 4.2 provides an overview of our data, while Section 4.3 outlines the forecasting methods used in our study. We also describe our methodology for data sampling in Section 4.4. Section 4.5 presents our results, and we conclude the chapter in the final section.

4.1 Literature review

Extensive research has been conducted on the factors influencing RRs. In this section, we begin by examining theoretical models that explain the factors that determine RRs, while also highlighting potential distinctions between corporate bonds and consumer credit. Subsequently, we investigate empirical evidence that quantifies the significance of macroeconomic predictors in predicting recovery rates, comparing the findings in the context of defaulted corporate bonds and consumer credit.

The recovery rate (RR) for corporate loans and bonds depends on the value stream that can be obtained from liquidating a firm's assets in the event of default. The expected residual value to claimants is greater when the firm's value is higher and its level of indebtedness is lower (Merton, 1974). The macroeconomic and financial conditions that firms and their peers operate in can affect the value stream generated by the firm's assets, as well as the revenues and costs associated with bankruptcy proceedings in the case of default, thus directly impacting the RR (Shleifer and Vishny, 1992). This idea has been supported by extensive empirical evidence, with studies such as Altman, Brady, Resti, and Sironi (2005); Acharya, Bharath, and Srinivasan (2007); Bruche and González-Aguado (2010); Gambetti, Gauthier, and Vrins (2019) highlighting the importance of macro-financial conditions in explaining the determinants of the recovery rate of corporate bonds.

Conversely, we only know little about consumer credit recovery rates, including whether it displays a pro-cyclical pattern similar to that observed in corporate bonds or consumer default rates (Nakajima and Ríos-Rull, 2014; Livshits, 2015). However, from a theoretical perspective, we can expect that the MS conditions should influence similarly the recovery rate for consumer credit. The income and residual value of consumers' personal assets are contingent on the environment in which the recovery process takes place. However, differences in access to information and legal frameworks make recovery procedures for consumer debt and corporations inherently different. While firms are legally obligated to periodically disclose their balance sheets and income statements publicly, accessing such information for final consumers and small and medium enterprises (SMEs) can be costly or even impossible. This poses a significant obstacle when pursuing credit recovery, especially in the case of unsecured consumer credit. The legal environment may also restrict access to information and limit actions to pursue recovery. For example, in the credit card industry, Fedaseyeu (2020) found that consumer protection legislation governing third-party debt collection reduces the number of third-party debt collectors and increases the loss given default on delinquent credit card loans.

In contrast to corporate bonds, the empirical evidence on the importance of MS indicators for explaining consumer credit RR is inconclusive. The results appear to vary

based on the time in which the studies were conducted, the specific type and sector of the NPC, and the number of MS variables that were considered.

For example, the study by Bellotti and Crook (2012) focused on defaulted credit-card accounts from 1999 to 2005 and found that including macroeconomic variables may enhance the predictive power of recovery rates for defaulted credit-card accounts. They only include three macroeconomic series and found that higher interest rates and unemployment levels at the time of default tend to result in lower recovery rates, while higher earnings growth leads to better recoveries. Conversely, Leow, Mues, and Thomas (2014) find that macroeconomic variables may improve the predictive performance of mortgage loan RR estimates, but not of personal loan RR. They consider a more complete set of economic variables at the UK national level, including net lending growth, disposable income growth, GDP growth, net lending growth on dwellings, unemployment rate, saving ratio, interest rate, and the Halifax House Price Index with data on defaulted mortgages and unsecured personal banking loans ranging from 1990 to 2002 and from 1989 to 1999, respectively.

Looking at the broader market of consumer credit, Thomas, Matuszyk, and Moore (2012) compared debt characteristics and the RR of consumer credit for in-house and third-party collection policies. However, their analysis does not consider macroeconomic and financial predictors and it only considers defaulted consumer credits over two years in the 1990s (for in-house collection) and 2000s (for third-party collection). Beck, Grunert, Neus, and Walter (2017) examined the RR of consumer credits for goods and services in the period 2004–2008 in Germany, but mainly focused on idiosyncratic determinants such as the EAD or prior debtor-specific collection rates. Only the growth rate of gross domestic product and the unemployment rate in the debtor's federal state for the year in which the account was handed over to a third-party collection agency were considered. They have mixed results on the relationship between macroeconomic conditions and RR, with only the unemployment rate having a consistently negative effect on the recovery rate. Nazemi, Rezazadeh, Fabozzi, and Höchstötter (2022) consider 65,535 defaulted unsecured consumer credits purchased between 2010 and 2013 from a German telecommunications company. To study the effects on the recovery rate related to the economic conditions, they include the unemployment rate at the district level and *excessive indebtedness rate*, which measures the percentage of adults with strong negative credit ratings in a specific district. However, Nazemi, Rezazadeh, Fabozzi, and Höchstötter (2022) find that including such variables does not improve the prediction accuracy.

Finally, as underlined by Kalotay and Altman (2016) and Nazemi, Baumann, and Fabozzi (2022), evidence in the relevance of MS and NPC-specific predictors can be faulted. Recovery rate modeling has mostly focused on in-sample *vs* out-of-sample splits to analyze recovery rates determinants (Gambetti, Gauthier, and Vrins, 2019) or discuss the best predictive strategies (Nazemi and Fabozzi, 2018; Nazemi, Heidenreich, and Fabozzi, 2018; Nazemi, Baumann, and Fabozzi, 2022; Gambetti, Roccazzella, and Vrins, 2022). When the data is randomly assigned to in-sample and out-of-sample sets, realizations of RRs used for training the models have likely occurred after the defaults of credits used for evaluating the models. The underlying assumption is that the data-generating process is time-invariant which, if not met, can potentially lead to a look-ahead bias in the predictions, therefore invalidating the findings.

Contrasting to previous studies, with an unprecedented database at hand, these limitations can be overcome. We consider a broader range of consumer debt, including banking loans, financing contracts, sales credits, and debts related to telecommunication

and utilities. We also consider a larger set of 13 MS variables to describe macroeconomic and social conditions. Moreover, by conducting an inter-temporal prediction exercise, which ensures that only sample points observed before the default event are used during the training process, we may dispel the doubt on whether the relevance of MS predictors found in a simple out-of-sample exercise is confirmed.

4.2 Data

To properly understand the data associated with the consumer credit business, it is important to highlight its specificity, underlining its main features compared to other types of credits. We then describe the data provided by the third-party collector agency and the set of economic indicators we use in our analysis.

4.2.1 Non-performing consumer credit (NPC) database

The database under study (called NPC DB hereafter) is private and has never been exploited in any publication so far. It counts 6,493,794 non-performing consumer credits (NPCs) from Italy. It offers a representative picture of the situation in the country: all the 20 regions and 101 out of 107 provinces are represented. It contains defaulted consumer credit that the third-party collector is entitled to recover for a maximum duration period of 365 days since receiving the mandate from the original lender.

Table 4.1 reports the descriptive statistics of the recovery rates across types of debtors, and sectors and conditional on Total to Recover (TtR) ranging from EUR 50 to EUR 10,000. Among the debtors, 78.10% are private individuals and a small fraction (about 0.15%) are professionals, the information is missing for the remaining credits. The defaulted credits originate from the utility (35.04%) and telecommunication sectors (48.03%), from financing (car loans, credit cards, leasings, 5.45%), banking (overdrafts and personal credit, 3.72%) and commercial credit (sales credit, 7.78%).

We observe that NPCs originating from the telecommunication industry (Telco) display the lowest average recovery rate of 19%, followed by the utility and commercial sectors, yielding 27% and 31% recovery rates, respectively. NPCs originating from the banking and financing industry display instead an average 48% and 47% recovery rate, respectively. Figure 4.1 confirms the usual bi-modal distribution of recovery rates, with no-recovery ($RR = 0\%$) and full recovery ($RR = 100\%$) being the two most likely outcomes of the collection process.

For each credit, we know the total amount to recover (TtR) that includes the principal at inception, any additional fees and interest, the maximum duration of the recovery process, and the principal to total to recover ratio, which measures the importance of interest payments and ancillary fees relative to the total amount to recover. Table 4.1 confirms previous evidence on the negative relationship between recovery rate and total amount to recover (Nazemi, Baumann, and Fabozzi, 2022).

We count four categories of debtors, male and female physical debtors, professionals, and unspecified. In the case of the debtor being a physical person, we also observe the age. For professional and unspecified debtors, the variable *age* is set to its median in the sample, i.e., 55 years old and the binary variable *professionals* is included. The data set includes information on the debtor's postal code and the date when the third-party agency was authorized to recover the defaulted credit. This information enables us to associate

Table 4.1: Recovery rate descriptive statistics of non-performing credits by debtors type, sectors, and conditional on Total to Recover (TtR) ranging from EUR 50 to EUR 10,000.

				Percentile				
Debtor	# obs	μ	σ	.10	.25	.50	.75	.90
Female	3,001,281	0.25	0.4	0	0	0	0.5	1
Male	2,070,420	0.25	0.4	0	0	0	.5	1
Professionals	9,930	0.35	0.46	0	0	0	1	1
Unspecified	1,412,163	0.27	0.41	0	0	0	.622	1
Sector	# obs	μ	σ	.10	.25	.50	.75	.90
Telco	3,118,743	0.19	0.37	0	0	0	0	1
Commercial	504,604	0.31	0.43	0	0	0	.874	1
Utility	2,275,620	0.27	0.42	0	0	0	.723	1
Banking	240,936	0.48	0.47 0	0	.396	1	1	
Financing	353,891	0.47	0.37	0	0	.479	.886	1
TtR	# obs	μ	σ	.10	.25	.50	.75	.90
50-100	764,869	0.42	0.48	0	0	0	1	1
100-250	1,933,739	0.31	0.44	0	0	0	.909	1
250-500	1,787,211	0.21	0.37	0	0	0	.323	1
500-1,000	1,128,216	0.18	0.35	0	0	0	.13	.96
1,000-2,500	621,764	0.16	0.33	0	0	0	0	.909
2,500-5,000	190,501	0.14	0.31	0	0	0	0	.777
5,000-10,000	67,494	0.15	0.32	0	0	0	.004	.855

each debtor with a particular province and region at a specific point in time, which allows us to examine how NPC-specific predictors vary across Italy.

To better explain the heterogeneity of the data across regions and highlight possible trends, we divide the regions into three macro-territories: northern, central, and southern. We follow the classification of the Italian National Institute of Statistics (ISTAT), which classifies the eight administrative regions of Aosta Valley, Piedmont, Liguria, Lombardy, Emilia-Romagna, Veneto, Friuli-Venezia Giulia, and Trentino-Alto Adige as the North, while Lazio, Marches, Tuscany, and Umbria are classified as the Center. The South includes the islands of Sicily and Sardinia, as well as Abruzzo, Apulia, Basilicata, Calabria, Campania, and Molise[3], the North consists of eight administrative regions: Aosta Valley, Piedmont, Liguria, Lombardy, Emilia-Romagna, Veneto, Friuli-Venezia Giulia, and Trentino-Alto Adige while the Center encompasses Lazio, Marches, Tuscany, and Umbria. Finally, the South included the islands (Sicily and Sardinia) and Abruzzo, Apulia, Basilicata, Calabria, Campania, and Molise.

The top-left side of Figure 4.2 displays the average recovery rates and total recovery values for consumer non-performing credits in various Italian provinces, while the top-right side shows the same information for different regions. Generally, the recovery rates for non-performing credits are on average higher in the northern and central-northern regions of Italy (28%) compared to the center (26%) and southern regions and islands (22%).

[3]Information is available at: `https://www.istat.it/en/archivio/227202`

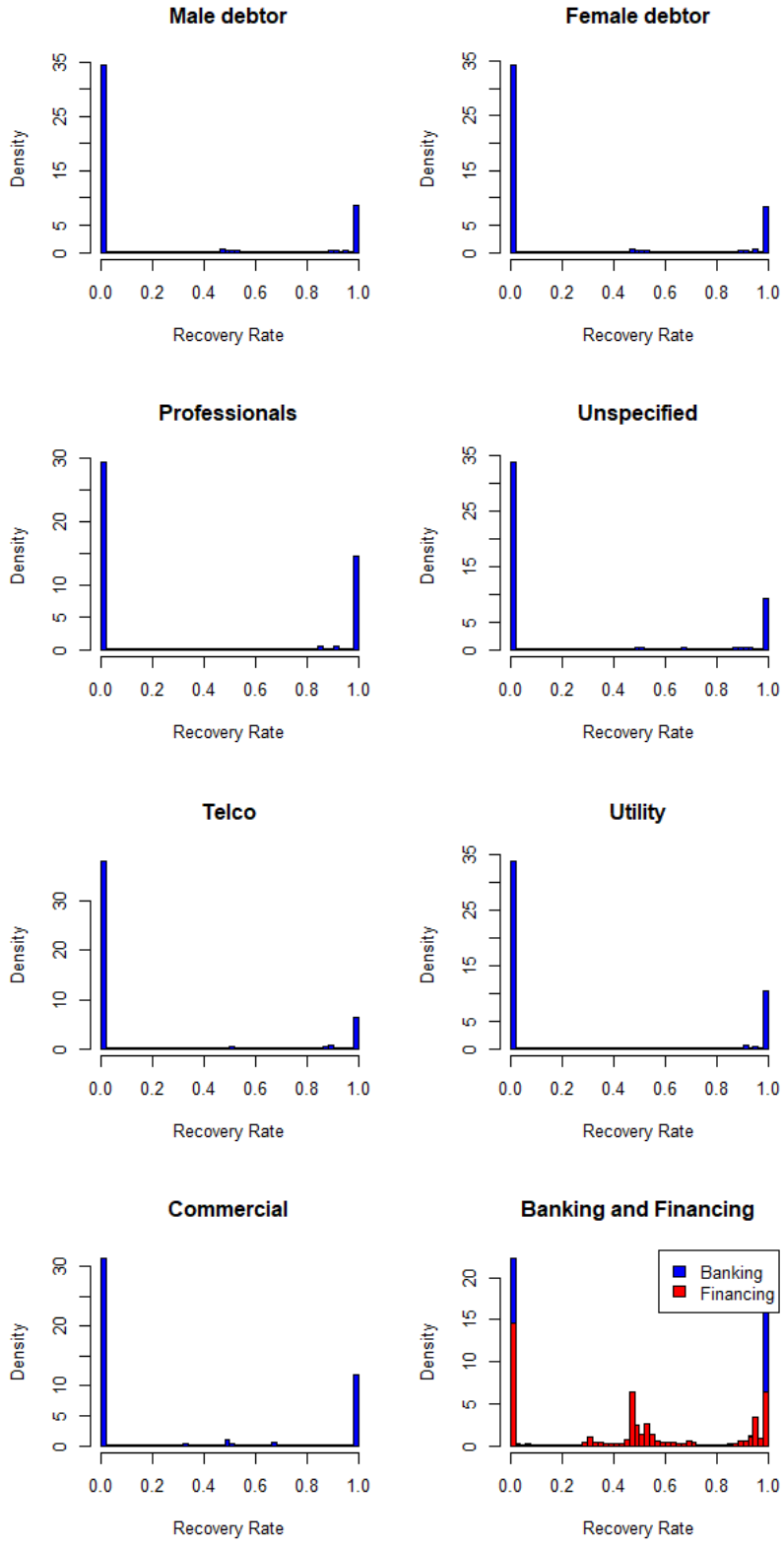


Figure 4.1: Distributions of the realized recovery rate of individual NPCs.

However, there is still a significant amount of variability at the provincial level. Contrasting to the trend in terms of recovery rates between the north and south, the bottom panel

of Figure 4.2 displays no consistent pattern for the total amount to recover, which is on average EUR 604 in the Center and North and EUR 550 in the South.

Figure 4.3 presents information on the average ratio of principal to the total amount to recover (top) and the average age of debtors (bottom). The data indicates that interest and ancillary fees have a relatively small impact on the total amount to be recovered, with the principal amount accounting for 86% to 96% of the total. This is particularly evident for NPCs originating from the center and south of Italy, where the principal accounts for over 90% of the total. Furthermore, debtors from the center and south of Italy have an average age of over 55 years, with some provinces such as Frosinone (Lazio), Caserta, Naples (Campania), and Messina (Sicily) having peaks of 60-66 years.

In the top panel of Figure 4.4, we can see the average maximum duration granted to third-party collectors to recover defaulted debt. The data shows that this duration is balanced across provinces and regions, with no significant trend emerging across the north, center, and south of Italy.

4.2.2 Economic and social predictors (collected outside of the NPCC DB)

We can now use the debtor's postal code and the date when the third-party agency was authorized to recover the defaulted credit to match each NPC to a specific region and province at a specific moment in time. This allows us to complete the set of predictors with macroeconomic and social indicators collected outside the NPC DB. We consider four broad classes of macroeconomic and social indicators measuring the effects of the business cycle, credit cycle, and economic uncertainty but also the social environment on RRs.

We start with the indicators of the business cycles. We consider the unemployment rate and the year-on-year growth rate of the Harmonized Index of Consumer Prices (HICP) as a proxy for inflation.[4] Unfortunately, because of structural breaks in the reporting of inflation and unemployment series at the regional or provincial level, we stick with these measures at the national level. We overcome this limitation when dealing with real gross value added as a measure of actual economic activity. This data is obtainable at the regional level and on an annual basis. We also incorporate its yearly growth rate as a gauge of business cycle fluctuations. Figure 4.6 illustrates the average real gross value added and yearly growth rate, highlighting significant economic differences between regions in terms of growth and levels.

The public debt and cost-of-debt to GDP ratios are used to evaluate credit conditions at both the national and regional levels. In Figure 4.5, the lower panel shows the variations in the cost-of-debt to GDP (on the left-hand side) and debt-to-GDP ratios (on the right-hand side) across different regions. Generally, the central and southern regions have a higher level of debt relative to their GDP and also experience a higher cost of debt compared to the northern regions. We also incorporate the spread between Italian and German 10-year constant maturity sovereign bond yields and the cost of debt to GDP ratio for the Italian economy to control credit market conditions at the national level. The economic policy uncertainty index of Baker, Bloom, and Davis (2016) for Italy is the last macroeconomic predictor we include.

The social environment can also significantly affect the outcome of the collection process. For instance, higher levels of criminal activity, lower income, or less developed areas may

[4]Both series are monthly and seasonally adjusted.

directly impact the collection process and affect the realized recovery rate. To capture the socio-economic heterogeneity at the province level, we examine four variables: 1) the number of robberies, 2) intentional homicides, 3) unintentional homicides per 100,000 inhabitants, and 4) the Life Quality Index published by the Italian financial newspaper Il Sole 24 Ore. This index consists of 90 indicators grouped into six macro-categories, namely wealth and consumption, business and work, environment and services, demographics and health, justice and security, and culture and leisure. Starting in 2019, each macro-category is made up of 15 sub-indicators, and provinces are awarded one thousand points for the best value and zero points for the worst in each sub-indicator. For each of the six macro-categories, a ranking is calculated based on the average score reported in the 15 sub-indicators, and the final ranking is based on the arithmetic mean of the six sector rankings. The lower panel of Figure 4.5 and Figure 4.6 reveal that Milan, Rome, and Naples provinces have higher numbers of robberies, and intentional and unintentional homicides per 100,000 inhabitants compared to other provinces. The upper panel of Figure 4.7 displays instead the variation in the Life Quality Index across regions and provinces of Italy. The data highlights that regions and provinces in the north and the center tend to have higher average values of the Life Quality Index (538 and 521, respectively), compared to the south (419.95).

Finally, we include fixed effect controls for the region of residence of the debtor and the industry from which the credit has originated, i.e., utilities, financing, banking, telecommunications, and commercials. We report the descriptive statistics of contract-specific and macroeconomic and social predictors in Table 4.2 and 4.3.

Table 4.2: List of considered macroeconomic, financial, and social (MFS) variables.

Variable	Abrev.	Mean	Std.Dev.	Min	Pctl. 25	Pctl. 75	Max
Regional Gross VA Mln	R-GVA	119414.493	96610.849	4103.1	64286.2	141279	344629.2
Regional Gross VA dln1	R-GVA d1	0.108	1.975	-8.569	-0.258	1.149	8.947
National Unemployment	Un	11.807	5.872	2.747	7.339	18.382	23.415
EPU Italy	EPU	110.972	32.992	31.702	89.521	130.025	241.018
HICP YoY	HCIP	0.861	0.825	-0.509	0.193	1.363	4.165
Regional Debt to GDP	R-DtGDP	0.07	0.029	0.024	0.042	0.09	0.149
NationalDebt to GDP	N-DtGDP	129.787	9.979	103.9	132.5	134.8	135.4
National Cost of Debt to GDP	N-CtGDP	4.148	0.484	3.6	3.8	4.6	5.2
Spread ITA-GER 10Y bonds	Spread	171.549	65.719	21.066	130.701	205.274	460.318
Life Quality Index	LQ	490.15	66.603	359.97	425	545.55	641.49
Intentional Homicides	INHom	11.03	16.52	0	1	13	109
Unintentional Homicides	UNHom	30.931	31.788	0	10	40	138
Robberies	Rob	1337.33	2202.379	4	91	1799	14045

4.2.3 Portfolios of defaulted consumer credits

Third-party collectors acquire *the right to recover* defaulted credit in lots by participating in auctions with information on the sector and the geographical area originating the pool of credits and the maximum duration granted to the collection process but very little information on the specific constituents of the lot. They handle the recovery process in the name of the lender and earn a commission on the total amount recovered and on the number of collection actions performed. Therefore, studying the determinants at the portfolio level better reflects the third-party collectors' business model.

We now describe how we formed the portfolios of defaulted consumer credits we use for our analysis. For each quarter between (and including) 2007 Q1 and 2019 Q4, we randomly

Table 4.3: List of considered loan-specific predictive variables, controls, and corresponding descriptive statistics after having performed the stratified sampling procedure to Banking, Commercial, Financing, Telecommunications, and Utilities as described in Section 4.4 .

Variable	Abrev.	Mean	Std.Dev.	Min	Pctl. 25	Pctl. 75	Max
Age	Age	55.335	17.572	21	43	67	100
Max Duration Collection	Max Dur.	90.877	76.28	0	30	120	365
Gender F	F	0.272	0.445	0			1
Gender M	M	0.425	0.494	0			1
Professionals	Prof.	0.006	0.076	0			1
Recovery Rate	RR	0.344	0.43	0	0	0.91	1
Total to Recover	TtR	727.48	1104.19	50.01	185.13	742.672	10,000
Principal to Total to Recover	PTtR	0.881	0.186	0	0.874	1	1
Banking	Bank.	0.2	0.4	0			1
Commercial	Comm	0.2	0.4	0			1
Financing	Fin.	0.2	0.4	0			1
Telecommunications	Telco	0.2	0.4	0			1
Utilities	Util.	0.2	0.4	0			1
Region: Abruzzo	ABR	0.026	0.158	0			1
Region: Basilicata	BAS	0.008	0.091	0			1
Region: Calabria	CAL	0.034	0.181	0			1
Region: Campania	CAM	0.145	0.352	0			1
Region: Emilia Romagna	EMR	0.055	0.227	0			1
Region: Friuli Venezia Giulia	FVG	0.012	0.111	0			1
Region: Liguria	LIG	0.027	0.161	0			1
Region: Lombardia	LOM	0.135	0.342	0			1
Region: Marche	MAR	0.017	0.131	0			1
Region: Molise	MOL	0.006	0.076	0			1
Region: Piemonte	PIE	0.067	0.25	0			1
Region: Puglia	PIG	0.045	0.207	0			1
Region: Sardegna	SAR	0.011	0.102	0			1
Region: Sicily	SIC	0.107	0.309	0			1
Region: Trentino Alto Adige	TAA	0.006	0.076	0			1
Region: Tuscany	TUS	0.055	0.228	0			1
Region: Umbria	UMB	0.012	0.107	0			1
Region: Val d'Aosta	VDA	0.09	0.287	0			1
Region: Veneto	VEN	0.052	0.222	0			1

draw (without replacement) samples of 100 individual credits that have been acquired in that quarter. We consider portfolios of 100 credits because it offers an acceptable compromise between keeping enough observations for the training (and testing) set and maintaining the portfolio large enough to reduce the relevance of individual credits at the portfolio level. For each quarter, we continue to form portfolios till we have less than 100 individual credits available for sampling.

These portfolios can be representative of a third-party collector that at the end of each quarter forms portfolios using 100 of the defaulted consumer credits acquired during the reference quarter. The total amount to recover and the recovered amount will simply be the sum of the respective amounts of the credits included in the portfolio. The recovery rate of the portfolio is defined as the weighted average

$$RR_w = \sum_i w_i RR_i \quad \text{where} \quad w_i = \frac{TtR_i}{\sum_j TtR_j} \ .$$

with TtR being the total to recover. The predictor set for the portfolios (contract-specific and macroeconomic, financial, and social predictors) is obtained by the weighted averages of the corresponding predictors, using the same weights.

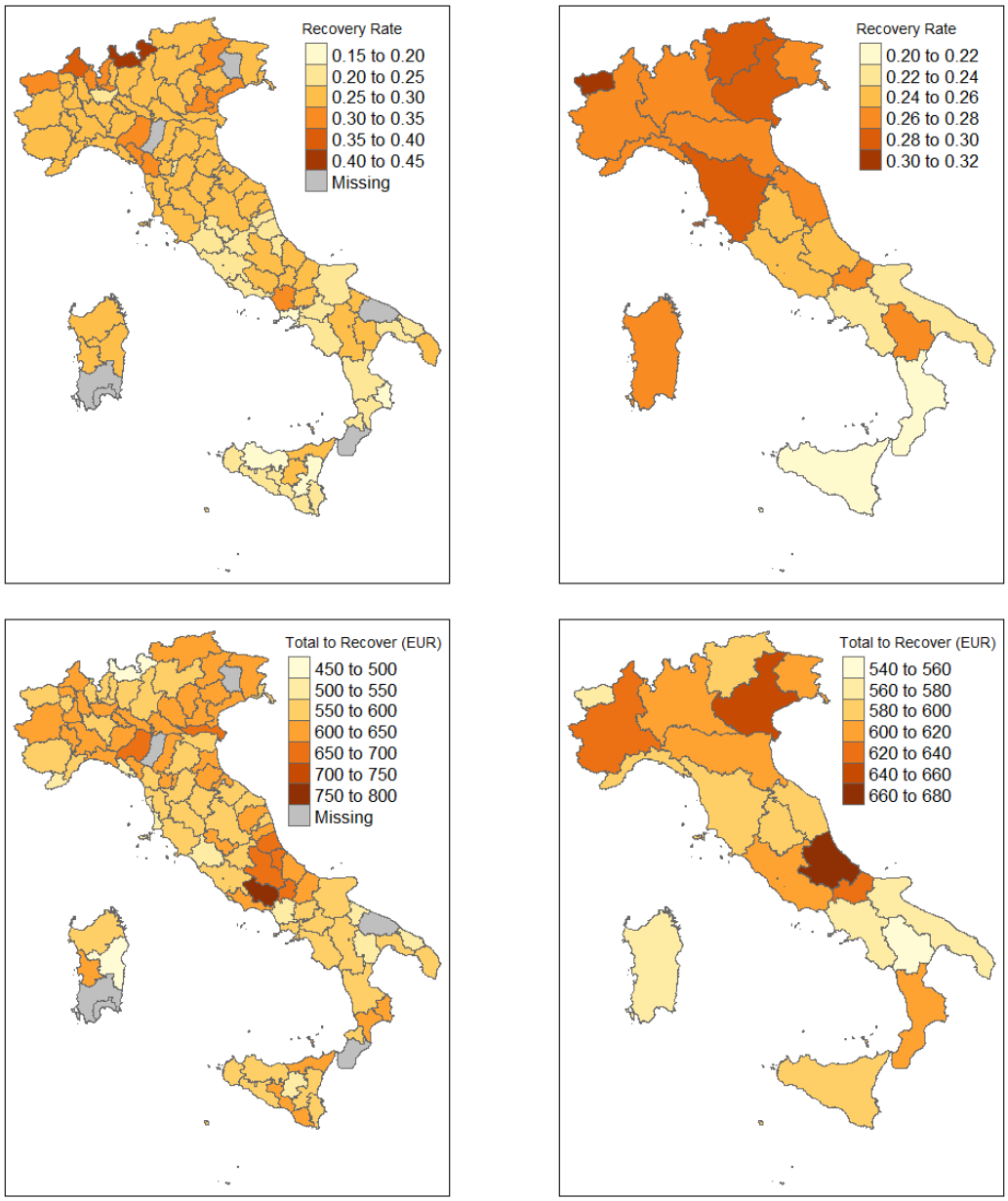


Figure 4.2: Heterogeneity of recovery rates and total to recover amounts at province and region level.

4.2.4 Inter-temporal and out of sample evaluation

The literature on recovery rates modeling has mostly focused on in-sample vs out-of-sample frameworks to analyze recovery rates determinants (Gambetti, Gauthier, and Vrins, 2019) or discuss the best predictive strategies for bonds and loans recovery rates Bellotti, Brigo, Gambetti, and Vrins (2021); Nazemi, Heidenreich, and Fabozzi (2018). Nevertheless, Kalotay and Altman (2016) and Nazemi, Baumann, and Fabozzi (2022) question the applicability of conventional out-of-sample evaluation in the field. Specifically, when the data is randomly partitioned into in-sample and out-of-sample sets, realizations of recovery rates used for training the models have likely occurred after the defaults of credits used for evaluating the models. The underlying assumption is that the data-generating

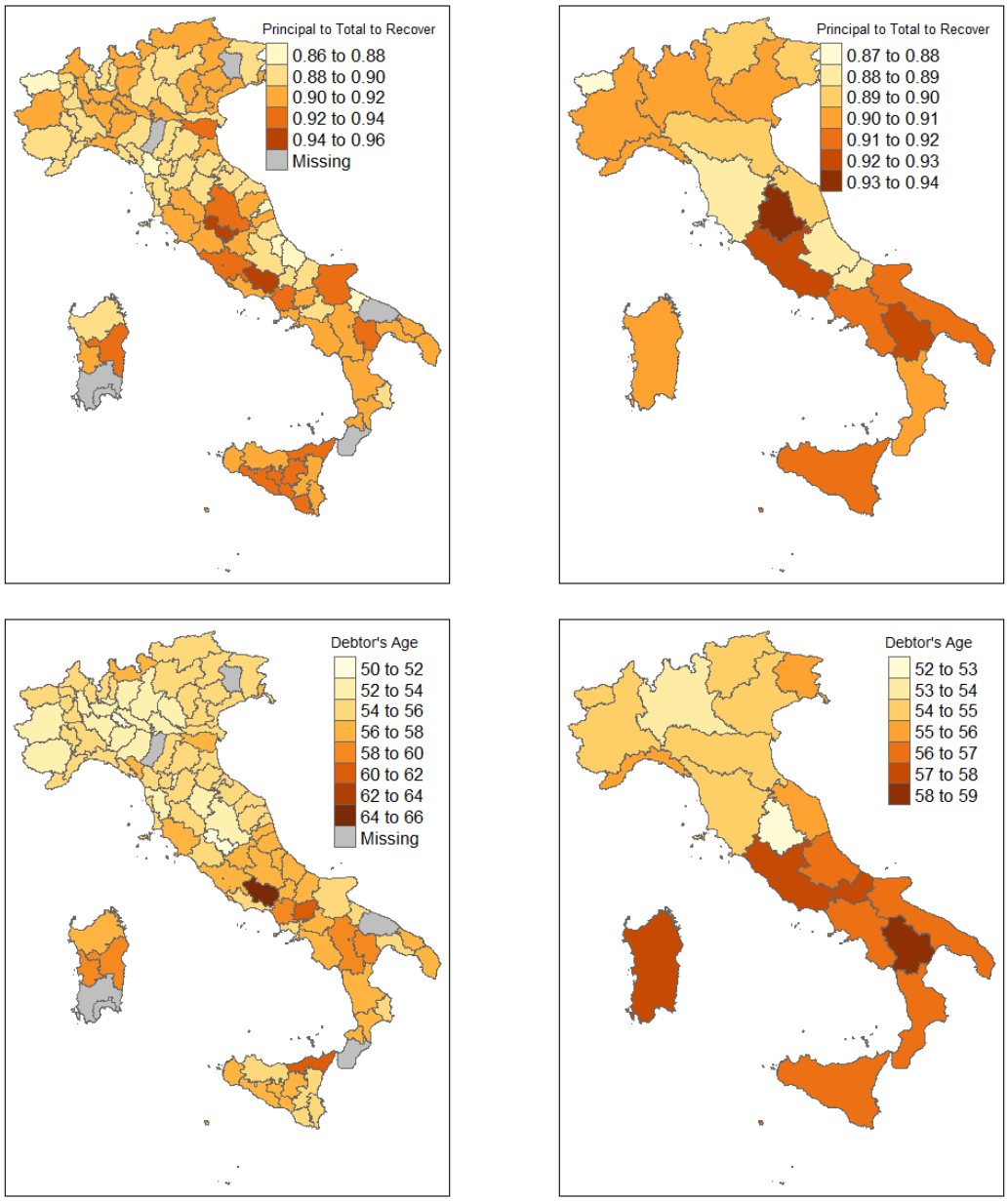


Figure 4.3: Heterogeneity of the principal to total to recover and of the debtors' age at province and region level.

process is time-invariant, which potentially leads to a look-ahead bias in the predictions. Moreover, in practice, third-party collectors can only use the information available at the time of acquisition of the defaulted credits to estimate their predictive model. This time inconsistency may pose some questions on the statistical appropriateness of standard in-sample vs out-of-sample analysis but also the practical relevance of the results for real-world users. To address this potential issue, we consider as a robustness check an *inter-temporal* prediction exercise by ensuring that only sample points observed before the default event were used during the training process.

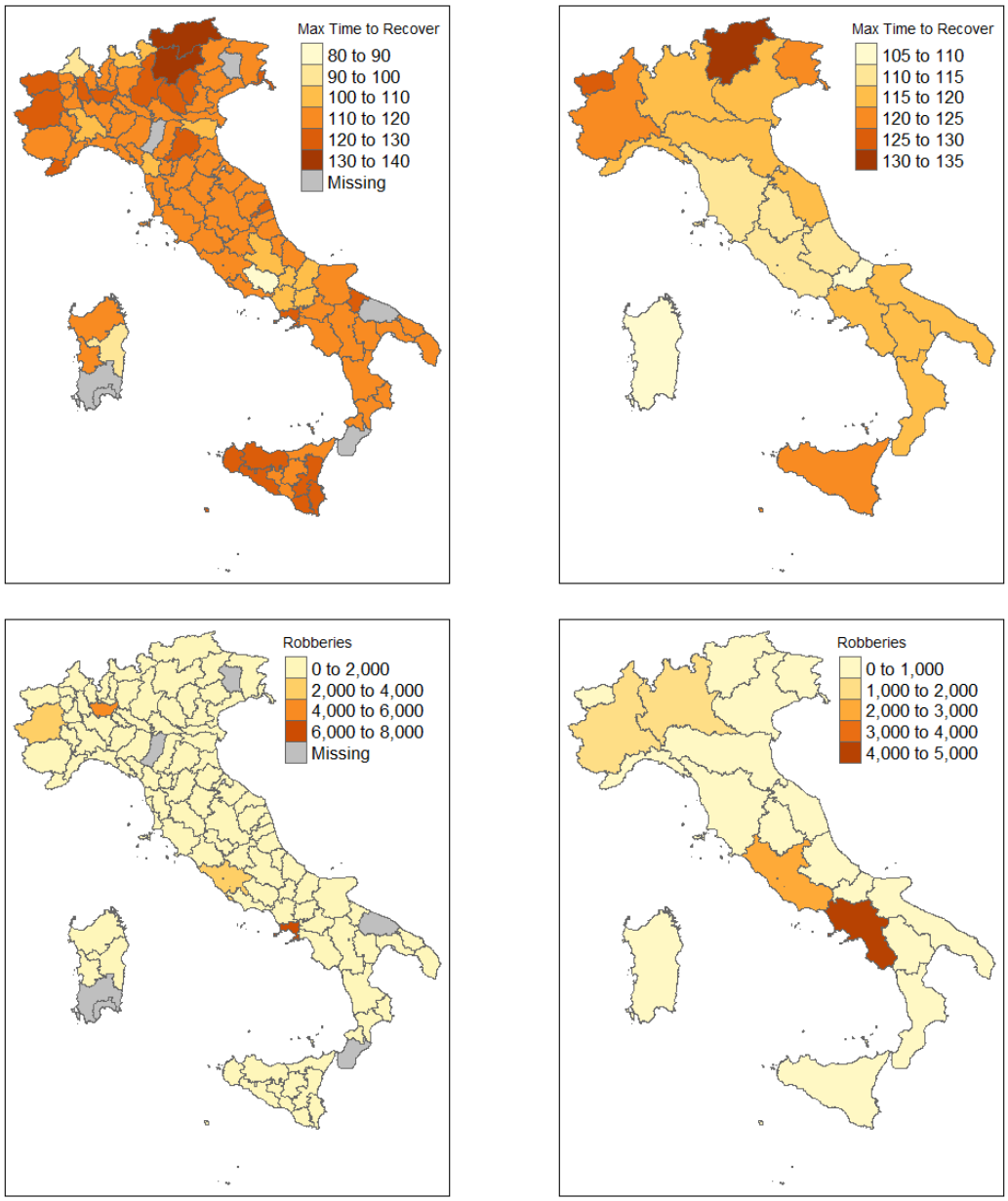


Figure 4.4: Heterogeneity of the maximum duration collection process and the number of robberies at province and region level.

4.3 Forecasting methods

Academic research highlights the potential of machine learning (ML) techniques in forecasting recovery rates (Qi and Zhao, 2011; Loterman, Brown, Martens, Mues, and Baesens, 2012; Yao, Crook, and Andreeva, 2017; Nazemi, Heidenreich, and Fabozzi, 2018; Nazemi and Fabozzi, 2018; Hurlin, Leymarie, and Patin, 2018; Nazemi, Pour, Heidenreich, and Fabozzi, 2017; Bellotti, Brigo, Gambetti, and Vrins, 2021; Nazemi, Rezazadeh, Fabozzi, and Höchstötter, 2022; Gambetti, Roccazzella, and Vrins, 2022), overcoming many of the limitations of the traditional multivariate linear and beta regressions used in earlier

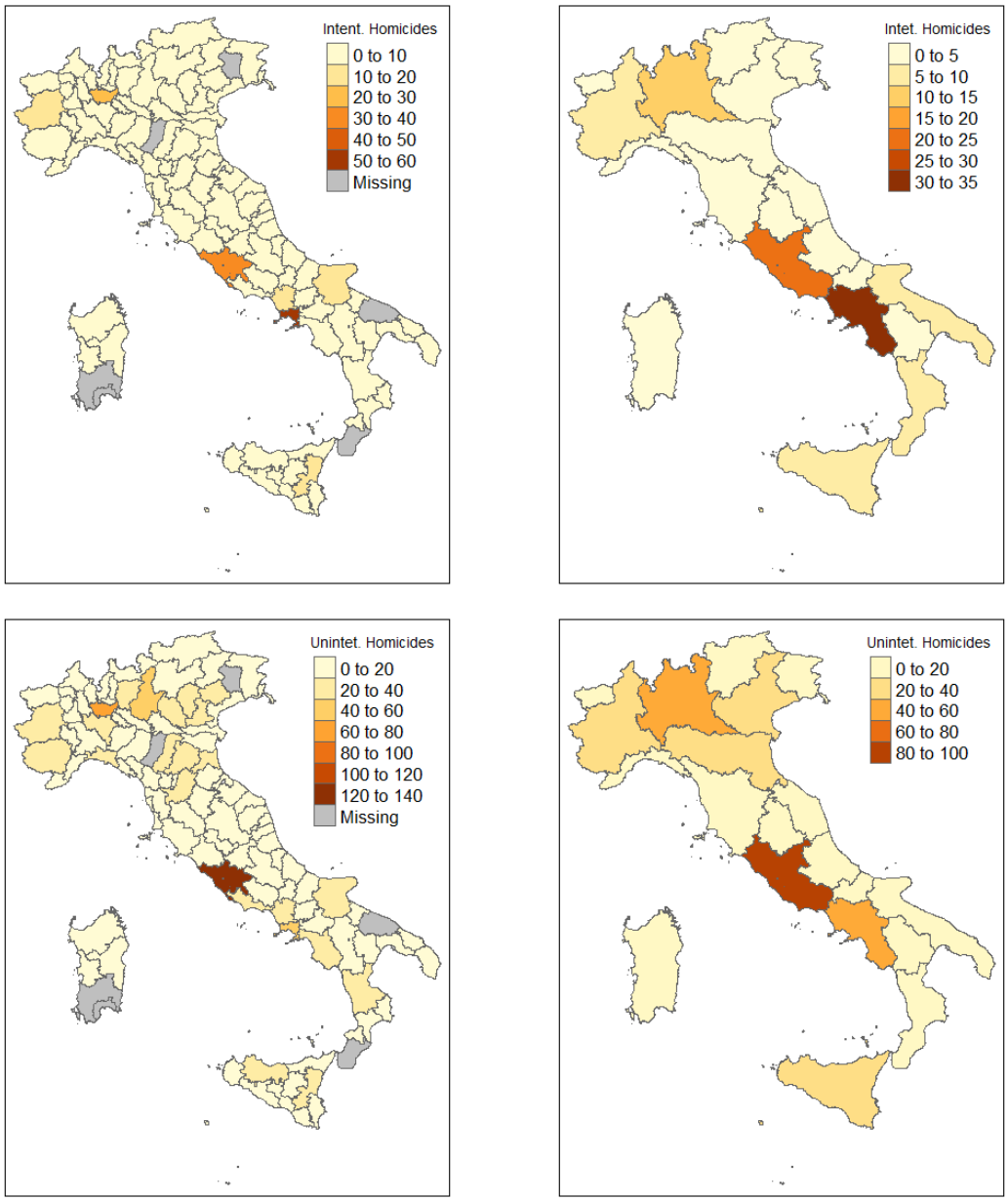


Figure 4.5: Heterogeneity of the number of intentional and unintentional homicides at province and region level.

studies.[5] The promising performance and the flexibility of machine learning, however, come at some costs. First, the estimated models are not always easy to interpret and it is not clear whether they also capture well-known economic and financial relationships. Second, it is not easy to identify the *right tool for the job* in the plethora of supervised ML algorithms. Third, machine learning methods can be computationally expensive. This

[5]For example, given that recovery rates are defined in the closed interval $[0, 1]$, predictions arising from a linear regression framework may lead to negative LGDs or values exceeding the unity, also invalidating inference based on Gaussian errors. Moreover, there is empirical evidence that the distribution of individual credit recovery rates is bi-modal, with no or full recovery as the most likely outcome of the collection process. As a result, the relationship between recovery rates and their determinants becomes inherently nonlinear.

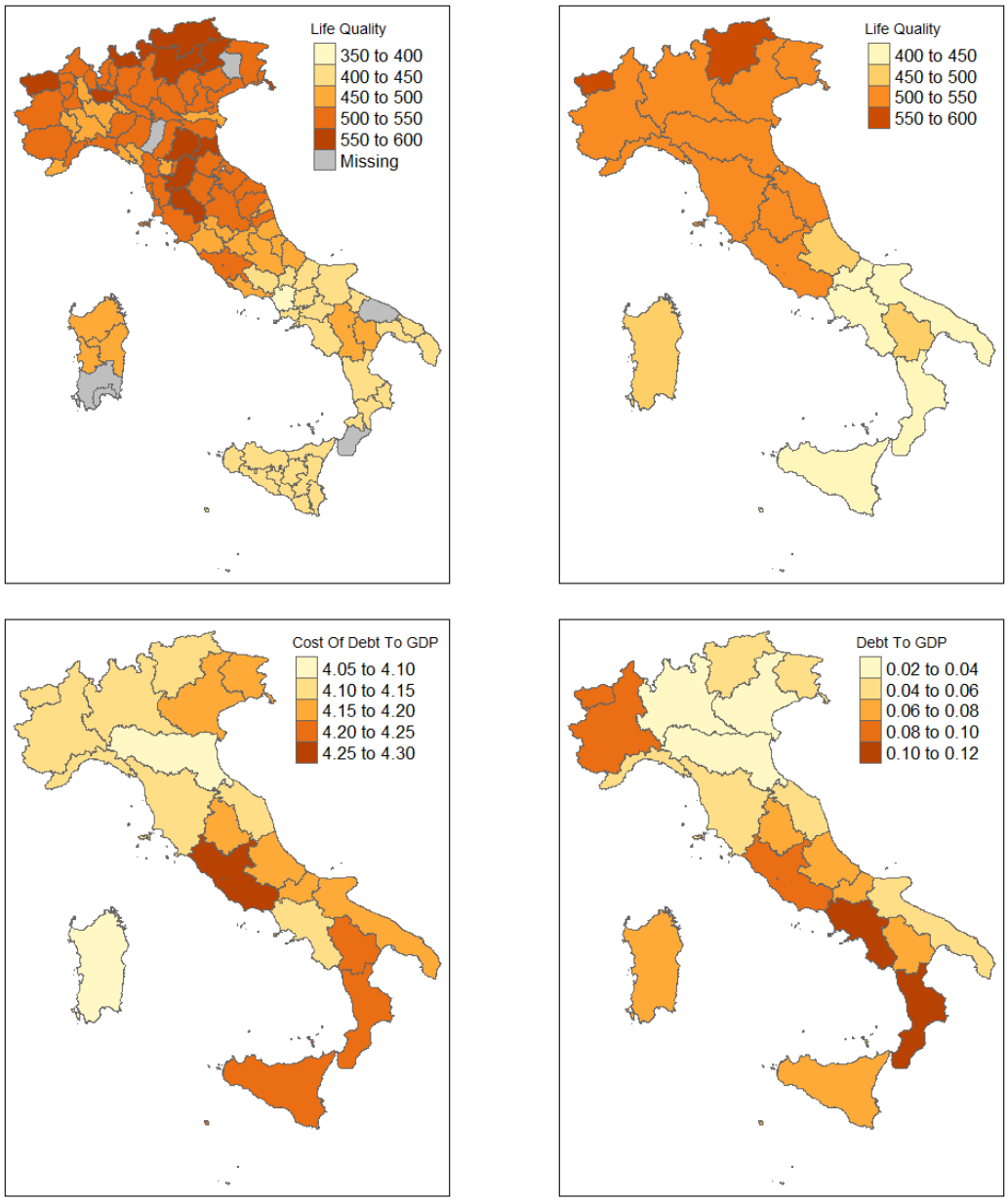


Figure 4.6: Heterogeneity of Life Quality, Cost of debt and Debt to GDP rations at province and region level.

last point is of particular concern in our study which counts millions of individual credits.

In this study, we employ well-known predictive models in credit scoring and recovery rate modeling literature (Loterman, Brown, Martens, Mues, and Baesens, 2012; Bellotti, Brigo, Gambetti, and Vrins, 2021). We include gradient-boosted regression trees, random forests, (bagged and individual) multivariate adaptive splines. We also consider the linear regression model estimated with ordinary least squares (OLS) and beta regression (Ferrari and Cribari-Neto, 2004) as benchmarks. Among these models, gradient-boosted regression trees, in particular, have proven to offer solid forecasting performance while remaining computationally tractable (Schmitt, 2022). We also explore the possibility of linear and nonlinear model combinations, which may offer diversification gains and are attractive when we cannot identify ex-ante the best single model or the employed predictive schemes

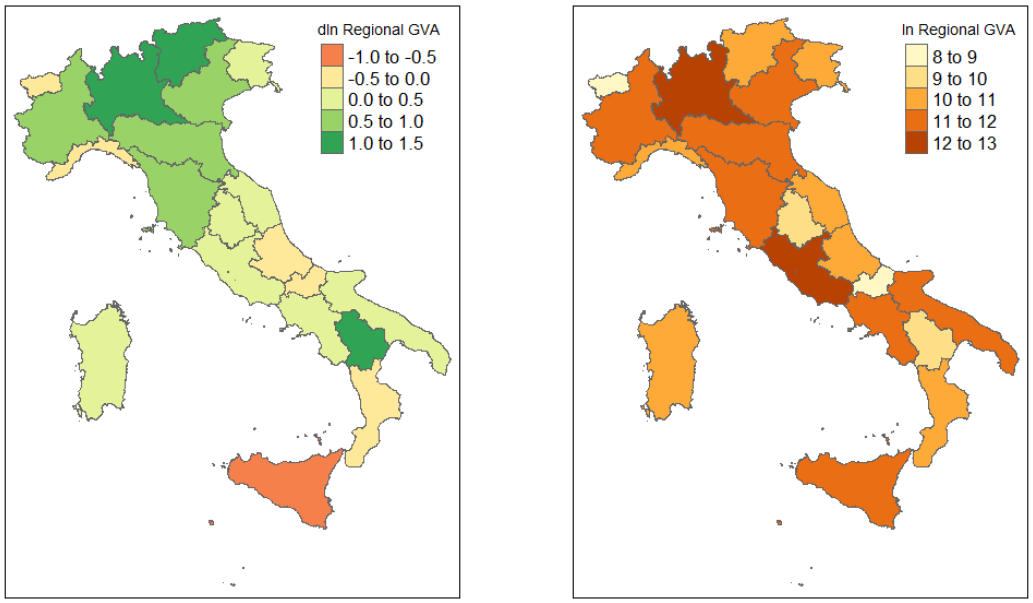


Figure 4.7: Heterogeneity of GDP level and growth at region level.

cover a wide spectrum of modeling assumptions (Atiya, 2020; Roccazzella, Gambetti, and Vrins, 2022b). We now proceed to briefly describe the predictive algorithms and the forecast combinations.

4.3.1 Linear and Beta regression models

Let p be the number of predictive variables and n be the sample size. We study the relationship between the n-dimensional vector of recovery rate $\boldsymbol{y}$ and the $n \times p$ matrix of predictive variables $\boldsymbol{X}$ in a beta regression model, which is tailored to model variables in the interval $(0, 1)$.[6] Following Ferrari and Cribari-Neto (2004), the beta regression model assumes that the density function of the dependent variable is a beta distribution given by

$$f(y, \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\phi\mu)\ \Gamma((1-\mu)\phi)} y^{\phi\mu-1}\ (1-y)^{(1-\mu)\phi-1}, \quad \text{with } 0 < y < 1\ , \tag{4.1}$$

with $E[y] = \mu$ and $Var[y] = \frac{\mu(1-\mu)}{1+\phi}$ and $\Gamma(\cdot)$ being the gamma function. Using this notation, the standard beta regression model of Ferrari and Cribari-Neto (2004) also assumes independent realizations of $y_i \sim \mathscr{B}(\mu_i, \phi_i)$, and a linear regression model for the transform of the mean parameter. In this chapter, we consider a *logit* link function for μ_i and a constant ϕ_i. This results in

$$\mu_i = \frac{e^{x_i^T \boldsymbol{\beta}}}{1 + e^{x_i^T \boldsymbol{\beta}}} \quad \text{and} \quad \phi_i = \phi\ . \tag{4.2}$$

Given that μ_i and ϕ_i are functions of the coefficients vector $\boldsymbol{\beta}$ and the scalar ϕ, respectively, we can estimate these parameters by maximum likelihood using a nonlinear optimization

[6] Following Smithson and Verkuilen (2006), we consider the transformation $\frac{\boldsymbol{y}(n-1)+0.5}{n}$ where n is the sample size to adapt the beta regression framework to dependent variables assuming also the extremes 0 and 1.

method. As a robustness check, we also consider the linear regression model,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u} \,. \tag{4.3}$$

This model can be easily estimated using OLS. However, RRs are defined in the closed interval $[0, 1]$, and therefore predictions arising from a linear regression framework may lead to recovery rates that are negative or greater than unity.

4.3.2 Gradient boosted regression trees and random forests

Tree-based methods can identify clusters of data with similar properties, to capture nonlinear relationships between the target and its predictive variables as well as to reproduce step-like relationships. Ensembles combine weak learners[7] to decrease the variance of a single prediction, therefore improving forecasting performance. *Bagging* and *boosting* are two popular types of ensemble methods.

Bagging (Breiman, 1996a) forms multiple versions of a model by making bootstrap replicates of the learning set and using these as new learning sets. Then, the aggregate prediction is formed by averaging the forecast over the versions. Random forests are an extension of bagged regression trees[8] However, bagged models may suffer from the risk of facing high correlations among individual models in the ensemble. Random forests (Breiman, 2001) is an ensemble of regression trees that additionally to standard *bagging*, also mitigates the risk of facing high correlations among individual trees in the ensemble by randomly selecting only a subset of predictors at each split of random forests. In this work, we consider an ensemble of 5,000 trees[9] and tune the number of randomly selected predictors via 5-fold cross-validation.

Boosting also uses multiple versions of the same models, but it differs from bagging in that the weak learners are trained independently. In boosting, models are trained sequentially and adaptively. *Adaptively* means that the second version of the model tries to correct the errors present in the first iteration. However, to avoid over-fitting, only a percentage of each fitted value (called *learning rate*) is subtracted from the residual from the previous learner. This procedure is continued and models are added until a stopping condition is met. The number of boosting iterations, i.e., the number of trees, the learning rate, and the individual trees' depth are the main hyper-parameters for this model. We employ the stochastic gradient boosting of Friedman (2002) that improves both the approximation accuracy and execution speed while mitigating the risk of overfitting. This is done by randomly drawing (without replacement) a sub-sample of the training data from the full training data set at each iteration and then using this one in place of the full sample to fit the base learner and compute the model update for the current iteration.

We also consider the XGBoost implementation of gradient-boosted trees (Chen and Guestrin, 2016). Similarly, to the stochastic gradient boosting of Friedman (2002), the number of trees, the learning (or shrinkage) rate, and the individual trees' depth are the main hyper-parameters. However, to further penalize complex model structures, mitigating the risk of overfitting, XGBoost uses L1 and L2 regularization on leaf weights.

[7] A weak learner is any machine learning algorithm that provides an accuracy slightly better than random guessing.

[8] Bagged regression trees are an ensemble learning method where multiple versions of regression trees are aggregated via bagging.

[9] Following Probst and Boulesteix (2017), we set this number to a computationally feasible large number, while still verifying that its cross-validated MSE does not exceed the ones obtained by smaller forests.

4.3.3 Multivariate adaptive splines

The multivariate adaptive regression splines algorithm (MARS) (Friedman, 1991) is a non-parametric regression model that considers piece-wise transformations of the original predictive variables. Specifically, given a set of cut points t and p predictors, each predictor x_j is transformed into a set of *reflected pairs* obtained by:

$$(x_j - t)_+ = \begin{cases} x_j - t, & \text{if } x_j > t \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad (t - x_j)_+ = \begin{cases} t - x_j, & \text{if } x_j > t \\ 0, & \text{otherwise} \end{cases} \tag{4.4}$$

where $j = (1, 2, ..., p)$ and $t \in \{x_{1j}, x_{2j}, ..., x_{nj}\}$. To determine the cut points, a linear regression model is estimated following a greedy procedure on the reflected pairs for each predictor. The reflected pairs that achieve the smallest mean square error are then used for the model. The number of selected pairs for each predictor defines the *degree* of the MARS algorithm. A backward *pruning* procedure handles over-fitting by dropping the features that are associated with the smallest error rate when excluded. In this study, we consider MARS with a degree 1, 2, and 3, while the number of terms to be retained in the final model can range between 5 and 100. The degree and pruning parameters are estimated via 5-fold cross-validation. We also consider a bagged version of the MARS algorithm. The ensemble is made of 30 versions of the MARS algorithm trained with a degree of 1 or 2, while the number of terms to be retained in the final model can range between 5 and 100. These parameters are tuned using 5-fold cross-validation.

4.3.4 Combinations of forecasts

We now describe methods to aggregate m individual forecasts into a single one. Formally, given a sample of n observations, let $\widehat{\boldsymbol{Y}} = \{\widehat{\boldsymbol{y}}_i, i = 1, ..., m\}$ be the $n \times m$ matrix containing the n forecasts produced by the m models, the vector of the aggregate forecasts is $\widehat{\boldsymbol{y}}_{FC} := \Phi(\widehat{\boldsymbol{Y}})$ using some aggregating function Φ.

In the special case of linear combinations, $\Phi(\widehat{\boldsymbol{Y}})$ takes the form of a weighted average $\widehat{\boldsymbol{Y}}\boldsymbol{w}$ where the weights w_i can be fixed (e.g., to $w_i = \frac{1}{m}$, leading to the simple average forecast) or estimated. A standard choice is to minimize the variance of the aggregate forecast error. In this case, assuming that the individual forecasts are unbiased, restricting the weights to sum to one keeps the aggregated forecast unbiased as well (Timmermann, 2006). With the sum-to-one and nonnegativity constraints on the combining weights, the optimization problem becomes

$$\boldsymbol{w}^* := \underset{\boldsymbol{w} \in \mathscr{W}}{\arg\min}\, \mathbb{E}\left\| \boldsymbol{y} - \widehat{\boldsymbol{Y}}\boldsymbol{w} \right\|_2^2 \quad \text{where} \quad \mathscr{W} := \{\boldsymbol{w} \in \mathbb{R}^m \mid \mathbf{1}^T\boldsymbol{w} = 1, \min \boldsymbol{w} \geq 0\}\ , \tag{4.5}$$

or, equivalently (Granger and Ramanathan, 1984):

$$\boldsymbol{w}^* := \underset{\boldsymbol{w} \in \mathscr{W}}{\arg\min}\, \boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w} \quad \text{where} \quad \mathscr{W} := \{\boldsymbol{w} \in \mathbb{R}^m \mid \mathbf{1}^T\boldsymbol{w} = 1, \min \boldsymbol{w} \geq 0\}\ , \tag{4.6}$$

where $\boldsymbol{\Sigma}$ is the population covariance matrix of models' prediction error, $\boldsymbol{v}_i := \boldsymbol{y} - \widehat{\boldsymbol{y}}_i$. This problem can be solved efficiently using quadratic programming or imposing the Karush-Kahn-Tucker conditions. Restricting the weights to be non-negative and to sum to one implicitly selects which forecasts to combine and also reduces the estimation error by

inducing a shrinkage-like effect on the covariance matrix of forecast errors (Jagannathan and Ma, 2003).

However, the population covariance matrix of the models' prediction error is unknown. We estimate $\boldsymbol{\Sigma}$ with its sample counterpart $\boldsymbol{S}$.[10]

We also explore the relevance of nonlinear weighting schemes for the forecasts combination. Specifically, we opt for averaged neural networks (NN_{FC}) that take the forecasts produced by RF, GBRT, XGBt, MARS, B-MARS, Beta, and OLS linear regression models as inputs and aggregate these into one prediction. NN_{FC} is an ensemble of one-hidden-layer feed-forward neural networks (Ripley, 1996) and with the *sigmoid* activation function. The ensemble counts 20 neural networks that are initialized by using different starting values for the parameters to estimate and that include a *decay* factor to penalize large coefficients.[11] These features moderate the risk of over-fitting.

Table 4.4: List of considered algorithms and corresponding `R` software.

Description	Acronym	`R` algorithm	Reference
Equally weighted average forecast	EW FC		
Optimal FC(+)	Opt_+	`quadprog`	
Optimal FC(+) with shrinkage to EW	COSE_+ FC	`quadprog`	Roccazzella, Gambetti, and Vrins (2022b)
Averaged neural networks	ANN FC	`avnnet`	Ripley (1996)
OLS Linear regression	LM	`lm`	
Beta regression	beta	`betareg`	Ferrari and Cribari-Neto (2004)
Multivariate adaptive regr. splines	MARS	`earth`	Friedman (1991)
Bagged MARS	B-MARS	`bagEarth`	Friedman (1991)
Random Forest	RF	`ranger`	Breiman (2001)
Gradient Boosted Regr. Trees	GBRT	`gbm`	Friedman (2002)
XGB Trees	XGBt	`xgboost`	Chen and Guestrin (2016)

Notes. The predictive algorithms can be retrieved via the `R` library `caret` (Kuhn, 2008) and refer to Kuhn and Johnson (2013) for a textbook treatment of the `R` packages. Hyper-parameters are tuned via 5-fold cross-validation. The shrinkage estimator of the covariance matrix of forecasting error is estimated using `covEstimation` (Ardia, Boudt, and Gagnon-Fleury, 2017)[12] and `FinCovRegularization` (Yan and Lin, 2016)[13] `R` packages. We refer to the Appendix for `R` pseudo-code.

4.4 Methodology

4.4.1 Data sampling for individual NPCs and portfolios

We perform our analysis at individual credit level using training and testing samples that are drawn without replacement using stratified sampling based on the origin of the defaulted credit, i.e., whether it relates to the telecommunication, utilities, banking, or consumer finance industry. This procedure balances the sample for the type of defaulted

[10] As a robustness check, we also consider the shrinkage estimator towards the diagonal matrix $\boldsymbol{S}_{shrink}$, which is equivalent to shrinking the optimal forecast towards the simple average prediction (Roccazzella, Gambetti, and Vrins, 2022b). Results are comparable and available upon request.

[11] Coefficients (called weights in neural networks) are prevented to become too large by penalizing their L2 norm similarly to the L2 penalty in the Ridge regression. The intensity of the penalization, called weight decay in neural networks, and the number of neurons are tuned via 5-fold cross-validation. We set the number of neural networks in the ensemble to 20 to limit the computational burden. Results remain comparable when also considering larger ensembles of 30 and 40 networks.

consumer credit and it is equivalent to the stratified sampling procedure based on seniority types commonly used for defaulted corporate bonds in Nazemi, Pour, Heidenreich, and Fabozzi (2017); Nazemi, Heidenreich, and Fabozzi (2018) and Gambetti, Roccazzella, and Vrins (2022). Therefore, the actual number of observations is reduced to 1,204,755 individual credits equally spread between the considered industries. We consider a 70% - 30% split between the training and testing sets.

We now describe how we formed the portfolios of defaulted consumer credits we use for our analysis. For each quarter between (and including) 2007 Q1 and 2019 Q4, we randomly draw (without replacement) samples of 100 individual credits that have been acquired in that quarter. For each quarter, we continue to form portfolios till we have less than 100 individual credits available for sampling. We consider portfolios of 100 credits because in some quarters we have at most 200 credits. This also offers an acceptable compromise between keeping enough observations for the training (and testing) set and maintaining the portfolio large enough to reduce the relevance of individual credits at the portfolio level. Figure 4.8 displays the number of portfolios over time.

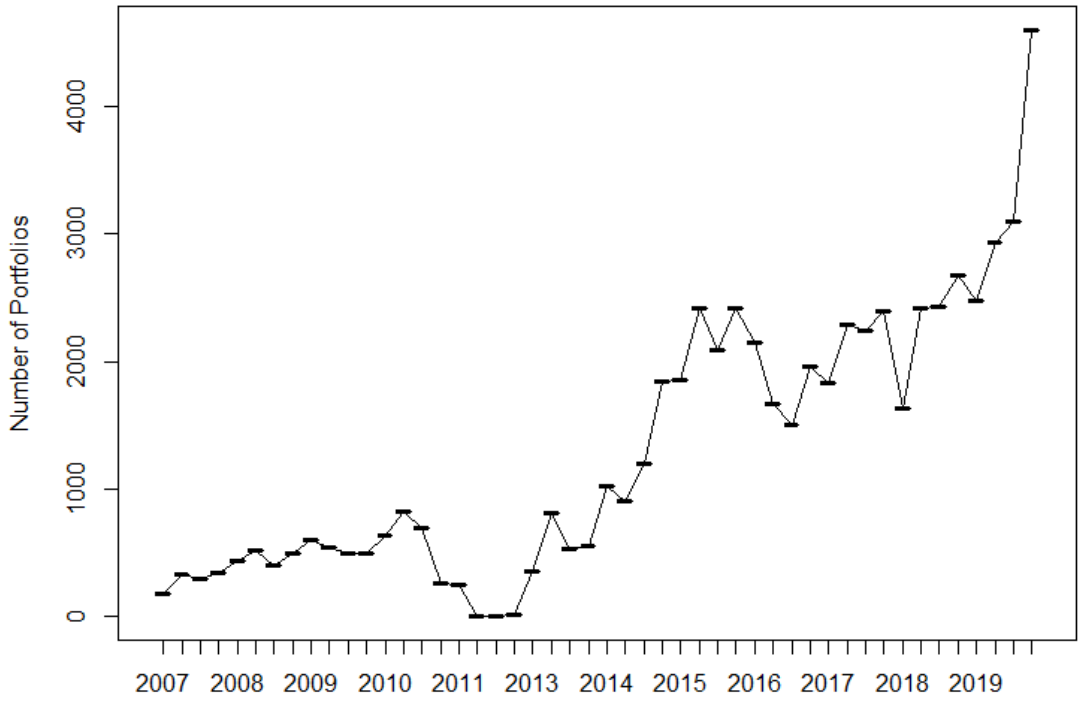


Figure 4.8: Number of portfolios over time.

The total amount to recover and the recovered amount will simply be the sum of the respective amounts of the credits included in the portfolio. The recovery rate of the portfolio is defined as the weighted average

$$RR_w = \sum_i w_i RR_i \quad \text{where} \quad w_i = \frac{TtR_i}{\sum_j TtR_j} .$$

with TtR being the total amount to recover. The predictor set for the portfolios (contract-specific and macroeconomic and social predictors) is obtained by the weighted averages of the corresponding predictors, using the same weights.

In total, we construct 62,082 portfolios and we consider a 70% - 30% split between training and testing sets. This implies that 43,457 observations are used for training while 18,625 were for the out-of-sample evaluation.

4.4.2 Time consistency: the inter-temporal exercise

Section 4.4.1 assumed that the data-generating process is time-invariant when randomly sampling data into in-sample and out-of-sample sets. We now explain how we design the *inter-temporal* prediction exercise that involved using only sample points that were observed before the default event during the training process.

Specifically, we use NPCs acquired before January 2017 for training, while credits acquired after January 2017 were used for testing. This corresponds to an approximately 50% - 50% split between training and inter-temporal testing sets permitting us to have enough observations to train data-intensive predictive models while keeping enough variability of yearly MS series.[14]

4.4.3 Evaluation criteria

We now describe the evaluation criteria to measure the potential benefit of including macroeconomic and social indicators in the predictors set while highlighting the differences (if any) between predicting recovery rates at the portfolio and individual credit levels.

We consider the root mean square error (RMSE), the mean absolute error (MAE), and the R^2. For a forecasting strategy j,

$$RMSE_j = \sqrt{\frac{1}{n_{test}} \sum_{i}^{n_{test}} (y_i - \widehat{y}_{i,j})^2} \,,$$

$$MAE_j = \frac{1}{n_{test}} \sum_{i}^{n_{test}} |y_i - \widehat{y}_{i,j}| \,,$$

$$R^2_j = 1 - \frac{\sum_i^{n_{test}} (y_i - \widehat{y}_{i,j})^2}{\sum_i^{n_{test}} (y_i - \overline{y})^2} \,,$$

where n_{test} is the number of observations in the test set, y_i is the i^{th} realization of the recovery rate, $\widehat{y}_{i,j}$ is the i^{th} forecast of the recovery rate obtained following the strategy j and $\overline{y}$ is the average recovery rate in the training set. Lower values of RMSE and MAE are preferred. Notice that R^2 can be negative. This signals that the forecast method j performs worse than a strategy considering the average recovery rate in the training set as a forecast. Following Nazemi, Rezazadeh, Fabozzi, and Höchstötter (2022), we also

[14] As a robustness check, we performed this analysis using the NPCs acquired before January 2016 were used for training, while credits acquired after January 2016 were used for testing. Results are also robust to this specification and they are available upon request.

include the Theil inequality coefficient (TIC), which is defined as

$$TIC = \frac{\sqrt{\frac{1}{n_{test}} \sum_i \left(y_i - \hat{y}_{i,j}\right)^2}}{\sqrt{\frac{1}{N} \sum_i y_i^2} + \sqrt{\frac{1}{n_{test}} \sum_i \hat{y}_{i,j}^2}} .$$

TIC rearranges the mean square error as the sum of the average quadratic realized and the estimated recovery rate. This indicator is constrained $[0, 1]$ and a value closer to 0 indicates higher prediction accuracy.

We compute the averages of out-of-sample (and inter-temporal) performance measures for the models via bootstrap, considering samples of 100 observations and repeating the procedure 10,000 times. In the inter-temporal exercise, 100 samples are bootstrapped at year level to handle the different number of portfolios. As in Nazemi, Baumann, and Fabozzi (2022), we use the Mann–Whitney test to assess whether the differences when including (*baseline specification*) and excluding macroeconomic financial and social (*exc. MS specification*) indicators predictors are statistically significant.

To identify the best forecasting approach(es) overall, we consider the model confidence set (hereafter MCS, Hansen, Lunde, and Nason, 2011) with respect to the square loss function, which tests whether a subset of methods enters jointly in the superior set of models by repeatedly testing the null hypothesis of equal predictive performance with significance level α.[15]

4.5 Results

Next, we present our findings. First, we examine the performance of the out-of-sample forecasting exercise on individual NPCs. Then, we proceed to the out-of-sample and inter-temporal exercise on portfolios of NPCs. When dealing with portfolios, all forecasting models are re-estimated using predictive variables at the pool level, mimicking a scenario in which third-party collectors lack information about the particular components of the pool and only possess data at the aggregate portfolio level. Finally, we employ Accumulated Local Effects (ALE) plots (Apley and Zhu, 2020) to depict how individual explanatory variables impact the model predictions, on average. This enables us to determine whether linear and beta regressions, as well as machine learning predictive models, also capture theoretically sound relationships between recovery rates and predictive variables.

Table 4.5 reports the RMSE, the MAE, the R^2, the TIC, and the models joining the superior set of models with 5% significance level. Table 4.6 reports the percentage difference of the performance when excluding MS indicators from the predictor set and the corresponding significance level of the Mann–Whitney test.

4.5.1 Forecasting performance

Our attention now turns to the outcomes of our predictive exercise using the standard out-of-sample setup. As shown in panels a) and b) of Table 4.5, we observe a substantial

[15] Formally, let $\mathscr{M}_0$ be the set of all forecasting models (both individual candidates and forecasts combinations), and let $\mathscr{M}^*$ be the superior set of models. Formally, the MCS tests $H_0 : \mathbb{E}\left[d_{i,j}\right] = 0, \ \forall i, j$. If the null hypothesis is rejected, then the procedure eliminates the model with the greatest relative loss from the set $\mathscr{M}_0$. This procedure is sequentially repeated until the null hypothesis is not rejected at the chosen probability level α. For the MCS, we consider the square loss and compute the p-values via bootstrapping (5,000 replications).

enhancement in predictive accuracy when working with portfolios, regardless of whether we incorporated MS predictors or not. For instance, if we choose Opt+ FC when MS predictors are excluded, the R^2 value rises from 28.27% when examining individual loans to 32.4% in the context of portfolios. This is not surprising because, similarly to the aggregation of individual stocks into portfolios, the idiosyncratic risk of individual credits' recovery rates tends to be diversified away. Moreover, the distribution of realized recovery rates for individual credits strongly differs from the one in the case of portfolios.

Figure 4.9 displays how the aggregation into portfolios affects the recovery rates distributions: the cases of no or complete recovery (0 and 100%) are diluted in portfolios, offsetting the bi-modality typical of individual recovery rates distributions. This observation is consistent with the theoretical result that describes the true and asymptotic distributions of the averaged recovery rate on an equally-weighted homogeneous pool, where defaults are controlled by a one-factor Gaussian copula (for further information, refer to Appendix A4.1.4).

As an additional finding, Table 4.5 also suggests that forecast combination techniques provide the most effective approach for predicting recovery rates at the individual credit and portfolio levels. In particular, the use of artificial neural networks as a nonlinear combination strategy is particularly encouraging since it is consistently included in the superior group of models and it attains the lowest MAE and RMSE values when MS indicators are included and when dealing with individual credits, portfolios, and portfolios in the inter-temporal setting.

Our focus now shifts to the importance of macroeconomic and social indicators. Panels a) and b) of Table 4.6 display the percentage differences in performance measures when MS indicators are excluded from the predictor set in the out-of-sample exercise for individual credits and portfolios, respectively. Performance significantly deteriorates when macroeconomic and social indicators are eliminated from the analysis. This is evident when forecasting the recovery rates of defaulted individual credits (panel a), but it is even more apparent at the portfolio level (panel b). When examining individual loans, the maximum improvement in R^2 is approximately 3 percentage points, achieved by using XGBt. However, the maximum improvement in portfolios is more than twice that amount: including MS predictors assists RF in increasing its R^2 by more than 6 percentage points.

Next, we perform the inter-temporal forecasting exercise that explicitly considers the time dimension in the estimation and testing steps. Panel c) of Table 4.5 indicates that forecasting performance deteriorates compared to the out-of-sample exercise, supporting the findings of Kalotay and Altman (2016) and Nazemi, Baumann, and Fabozzi (2022). However, Table 4.6 also shows how relevant and positive the effect of including macroeconomic and social indicators is. Specifically, all models improve their TIC; nine out of ten models improve their RMSE and R^2, and eight out of ten also improve the MAE. The size of this improvement is particularly noteworthy. For instance, the ANN FC model sees a rise of over 7 percentage points in its R^2, while the beta and linear regression models exhibit an increase of more than 15 percentage points.

4.5.2 Opening the black box using ALE plots

We have previously found that excluding macroeconomic and social indicators harms forecasting performance. Now, we focus on how economic as well as contract-specific variables influence the prediction of the models on average. To achieve this, we utilize

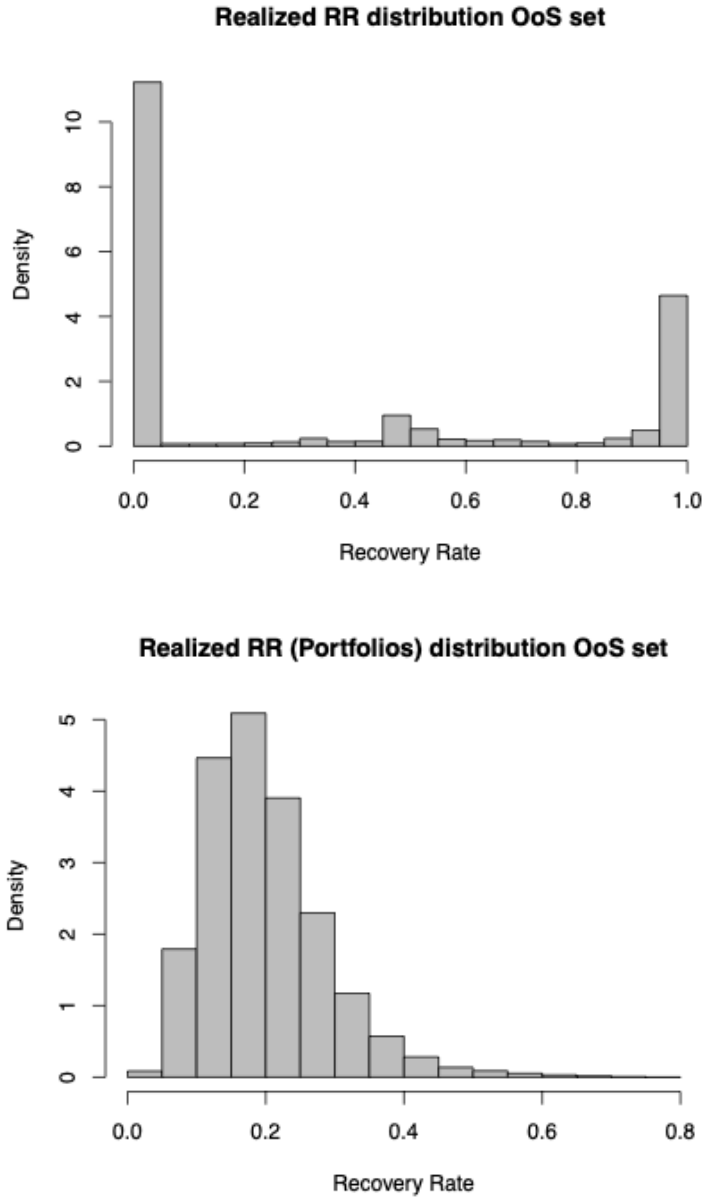


Figure 4.9: Distributions of the realized recovery rate for defaulted individual credits and portfolios of consumer credits.

Accumulated Local Effects (ALE) plots (Apley and Zhu, 2020), which show the changes in model predictions within a small neighborhood of the predictor of interest for data instances in that neighborhood. [16] We re-estimate the models by extending the training set to the full sample at the portfolio level to more precisely estimate ALEs for the entire range of values of the considered predictors. We selected the ALE plots in Figure 4.10, Figure 4.11, and Figure 4.12 based on their interpretability and the economic significance

[16] Another visualization technique is Partial Dependence (PD) plots, but they tend to be more computationally expensive and may produce inaccurate results when the predictors are highly correlated, as they require extrapolation of the prediction to values of the explanatory variable that are well outside the multivariate envelope of the training data. For a more in-depth discussion of ALE, PD, and marginal plots, we refer readers to Apley and Zhu (2020).

of the effect. The vertical axes measure the differences for the average prediction, while the horizontal axes the domain of the corresponding variable.

We start studying whether the industry originating the credit affects average recovery rates. ALE plots confirm what we have already observed in the summary statistics in Table 4.1. Setting the utility sector as a baseline, in *Banking* (Figure 4.10), we observe that models predict greater recovery rates when the proportion of credits from the banking sector in the portfolio increases. We observe a similar but less accentuated effect related to the quota of credits from the financing industry (*Financing*) and commercial credits (*Commercial*), but, conversely, the greater the proportion of defaulted credits from the telecommunication sector in the portfolio, the lower the expected recovery rate. This finding questions whether studies examining recovery rates for a specific sector can be generalized in the consumer credit industry, as previous research suggests. For instance, the greater expected recovery rate in *Banking*, may reflect the better screening ability of credit applicants by banking institutions. The specific types of consumer credit involved, i.e., overdrafts and personal credits, may also play an important role.

In the case of *Credit to Recover* (Figure 4.10), we confirm the negative correlation between the (log of the) total amount to recover and the recovery rate, as previously identified in the summary statistics presented in Table 4.1. This negative relationship has also been observed in defaulted consumer credits in the telecommunications industry (Nazemi, Baumann, and Fabozzi, 2022) and the corporate bond market (Bellotti, Brigo, Gambetti, and Vrins, 2021). Additionally, *Principal to Total to Recover* (Figure 4.10) indicates that a greater proportion of interest and ancillary fees in the total amount to recover leads to a lower expected recovery rate.

Moving to Figure 4.11, we observe that having been granted more time to perform the collection does not imply greater expected recovery rates (*Max Time to Recover*). Indeed, lenders can accept granting the third-party collector more time for the collection process if they expect the credit to recover to be more challenging. Setting Lazio region as a baseline, the region where the debtor resides may still affect expected recovery rates, showing that there are still many region-specific factors explaining recovery rates beyond the ones included in this study. Linear and beta regressions also capture a negative relationship between our proxies of criminal activity, i.e., the number of robberies and intentional homicides per 100.000 inhabitants, and recovery rates. However, the same variables seem to be irrelevant when looking at machine learning techniques.

In Figure 4.12, *Cost of Debt to GDP* helps us to study whether the dynamics of the credit cycle at the national level influence recovery rates for consumer credits. *Cost of Debt to GDP* signals how a higher level of interest to GDP at the Italian level can be associated with lower expected recovery rates. Similarly, we notice that only when the Italian debt to GDP ratio goes above 135%, recovery rates are expected to slightly decrease.

Figure 4.12 allows us to examine whether the national-level credit cycle dynamics affect recovery rates for consumer credit by focusing on *Cost of Debt to GDP*. We observe that higher levels of public debt interest in Italy are associated with lower expected recovery rates. Additionally, we notice that recovery rates are expected to slightly decrease when the *National Debt to GDP* ratio exceeds 135%. When looking at *EPU*, i.e., the ALE plot associated with the Economic Policy Uncertainty index of Beck, Grunert, Neus, and Walter (2017), we highlight how low uncertainty can be associated with higher expected recovery rates. Among the MS variables, the *EPU* has the largest effect on expected recovery rates, extending the findings of Gambetti, Gauthier, and Vrins (2019) to defaulted consumer

credits.

The accumulated local effect associated with linear and beta regression models underlines the negative relationship between *Unemployment* and expected recovery rate, however, machine learning algorithms seem not to be too sensitive to this predictor. Inflation at the national level, i.e., *HICP YoY*, seems positively associated with higher expected recovery rates, however, the estimated accumulated local effect varies across models. Concerning *HICP YoY*, however, we remark that Italy suffered from a deflationary trend following the outburst of the European sovereign bonds crisis in 2012. Therefore, it is not surprising that higher levels of inflation are also related to higher expected recovery rates. Nevertheless, we underline that the inflation level never went above 4% in the sample under analysis and therefore, we cannot assess the effect of high inflation rates on the recovery rate. Finally, a high level of *Regional GVA*, i.e., real output at the regional level, is also associated with a greater expected recovery rate, but the relationship is highly nonlinear, the size of the effect is only modest and it is only captured by gradient boosted trees.

4.6 Conclusions

We investigate the determinants of recovery rates in the consumer credit industry for credits originating from the utilities, banking, financial and commercial sectors. We also acknowledge that studying the determinants at the portfolio level is crucial to closely match the business model of third-party collectors that perform the recovery process in this industry.

Contrasting to previous studies, we find that macroeconomic and social indicators matter when forecasting recovery rates of defaulted consumer credits both at individual credit and (especially) portfolio level. This result is evident when looking at the sharp and significant decrease in forecasting performance of linear and beta regression models but also more advanced, yet standard, machine learning and forecast combination methods.

The analysis of accumulated local effects shows that the relationship between expected recovery rates and predictors related to the dynamics of the business cycle is also rather intuitive: a deterioration of the credit conditions, signaled by lower real activity, increasing cost-to-GDP ratio, and increasing economic uncertainty, is associated to lower expected recovery rates.

Table 4.5: Performance for the out-of-sample forecasting exercise at individual consumer loan and portfolio level for both specifications of the predictors set, i.e., when including or excluding MS. We highlight in bold the best performance measures for each specification. We report the models that have been included in the superior set of models at a 5% significance level in both MCS - incl. MS and MCS - excl. MS, based on whether they were estimated with or without MS predictors.

	RMSE	MAE	R2	TIC	RMSE	MAE	R2	TIC
	Predictors set excluding MS				Predictors set including MS			
a) Out of sample - Individual loans								
MARS	.3840	.3213	.1980	.4082	.3814	.3176	.2089	.4045
GBRT	.3676	.3004	.2646	.3864	.3632	.2952	.2818	.3809
XGBt	.3736	.3040	.2396	.3811	.3661	.2868	.2695	.3717
RF	.3673	**.2915**	.2654	.3805	.3615	.2867	.2880	.3730
B-MARS	.3847	.3249	.1952	.4112	.3822	.3210	.2054	.4071
Beta	.4282	.3990	.0046	.4576	.4247	.3946	.0205	.4528
OLS	.4027	.3473	.1189	.4370	.3987	.3422	.1364	.4306
Opt+ FC	**.3630**	.2960	**.2827**	**.3789**	.3567	.2884	.3074	**.3705**
EW FC	.3734	.3241	.2422	.3997	.3684	.3180	.2624	.3931
ANN FC	.3633	.2935	.2814	.3816	**.3561**	**.2801**	**.3093**	.3712
MCS L2 - incl. MS	ANN FC							
MCS L2 - excl. MS	·							
b) Out of sample - Portfolios								
MARS	.0745	.0556	.2820	.1759	.0725	.0540	.3198	.1708
GBRT	.0733	.0548	.3050	.1728	.0705	.0528	.3570	.1659
XGBt	.0740	.0559	.2912	.1726	.0718	.0542	.3314	.1672
RF	.0734	.0553	.3026	.1727	.0700	.0529	.3652	.1645
B-MARS	.0738	.0552	.2974	.1742	.0736	.0548	.3022	.1737
Beta	.0790	.0583	.2013	.1869	.0768	.0567	.2438	.1815
OLS	.0790	.0586	.1998	.1872	.0768	.0571	.2421	.1818
EW FC	.0733	.0551	.3078	.1731	.0711	.0535	.3489	.1676
Opt+ FC	**.0723**	**.0545**	.3240	.1703	**.0697**	.0526	**.3715**	.1639
ANN FC	.0724	.0546	.3205	**.1697**	**.0697**	**.0524**	.3690	**.1629**
MCS L2 - incl. MS	Opt+ FC ; COSE+ FC ; ANN FC							
MCS L2 - excl. MS	·							
c) Intertemporal - Portofolios								
MARS	.0813	.0592	.1590	.1894	.0772	.0589	.2414	.1663
GBRT	.0798	.0584	.1919	.1860	.0760	.0562	.2687	.1706
XGBt	.0852	.0636	.0802	.2053	.0809	.0607	.1712	.1884
RF	**.0780**	**.0576**	**.2280**	**.1813**	.0785	.0617	.2153	.1690
B-MARS	.0820	.0601	.1454	.1940	.0767	.0577	.2524	.1683
Beta	.0848	.0632	.0891	.2032	.0774	.0582	.2396	.1754
OLS	.0868	.0656	.0434	.2090	.0788	.0598	.2119	.1787
EW FC	.0811	.0601	.1665	.1921	.0758	.0572	.2721	.1690
Opt+ FC	.0784	.0578	.2199	.1829	.0764	.0590	.2580	.1671
ANN FC	.0793	.0584	.2015	.1860	**.0753**	**.0572**	**.2796**	**.1651**
MCS L2 - incl. MS	ANN FC							
MCS L2 - excl. MS	·							

Table 4.6: Loss differentials for the out-of-sample forecasting exercise when excluding macroeconomic, financial, and social indicators.

	RMSE		MAE		R2		TIC	
a) Out-of-sample - Individual loans								
MARS	0.69%		1.17%		-1.08		0.91%	
GBRT	1.20%		1.76%	*	-1.72	*	1.44%	
XGBt	2.06%	**	6.00%	***	-2.99	**	2.52%	**
RF	1.58%		1.67%		-2.26	*	2.01%	*
B-MARS	0.65%		1.21%		-1.02		1.01%	
Beta	0.82%	*	1.13%	***	-1.59	***	1.05%	***
OLS	1.01%	*	1.49%	**	-1.74	**	1.49%	***
EW FC	1.37%	*	1.92%	**	-2.03	**	1.68%	**
Opt+ FC	1.78%	*	2.63%	**	-2.47	**	2.26%	**
ANN FC	2.03%	*	4.80%	***	-2.79	**	2.81%	**
b) Out-of-sample - Portfolios								
MARS	2.76%	**	2.96%	**	-3.78	.014	2.99%	*
GBRT	3.97%	**	3.79%	**	-5.20	.006	4.16%	**
XGBt	3.06%		3.14%	**	-4.02	.028	3.23%	**
RF	4.86%	**	4.54%	***	-6.26	.000	4.98%	***
B-MARS	0.27%		0.73%		-0.48	.570	0.29%	
Beta	2.86%	*	2.82%	*	-4.25	.000	2.98%	**
OLS	2.86%	*	2.63%	*	-4.23	.000	2.97%	**
EW FC	3.09%	*	2.99%	**	-4.11	.004	3.28%	**
Opt+ FC	3.73%	*	3.61%	**	-4.75	.004	3.90%	**
ANN FC	3.87%	*	4.20%	***	-4.85	.016	4.17%	**
c) Intertemporal - Portfolios								
MARS	5.31%	***	0.54%		-8.24	***	13.91%	***
GBRT	5.05%	***	3.93%	**	-7.68	***	9.01%	***
XGBt	5.33%	***	4.77%	***	-9.09	***	8.96%	***
RF	-0.60%		-6.66%	***	1.27		7.27%	***
B-MARS	6.98%	***	4.24%	***	-10.70	***	15.24%	***
Beta	9.52%	***	8.65%	***	-15.05	***	15.86%	***
OLS	10.25%	***	9.78%	***	-16.86	***	16.98%	***
EW FC	7.04%	***	5.19%	***	-10.56	***	13.66%	***
Opt+ FC	2.62%	*	-2.05%	**	-3.80	***	9.44%	***
ANN FC	5.36%	***	2.03%		-7.81	***	12.62%	***

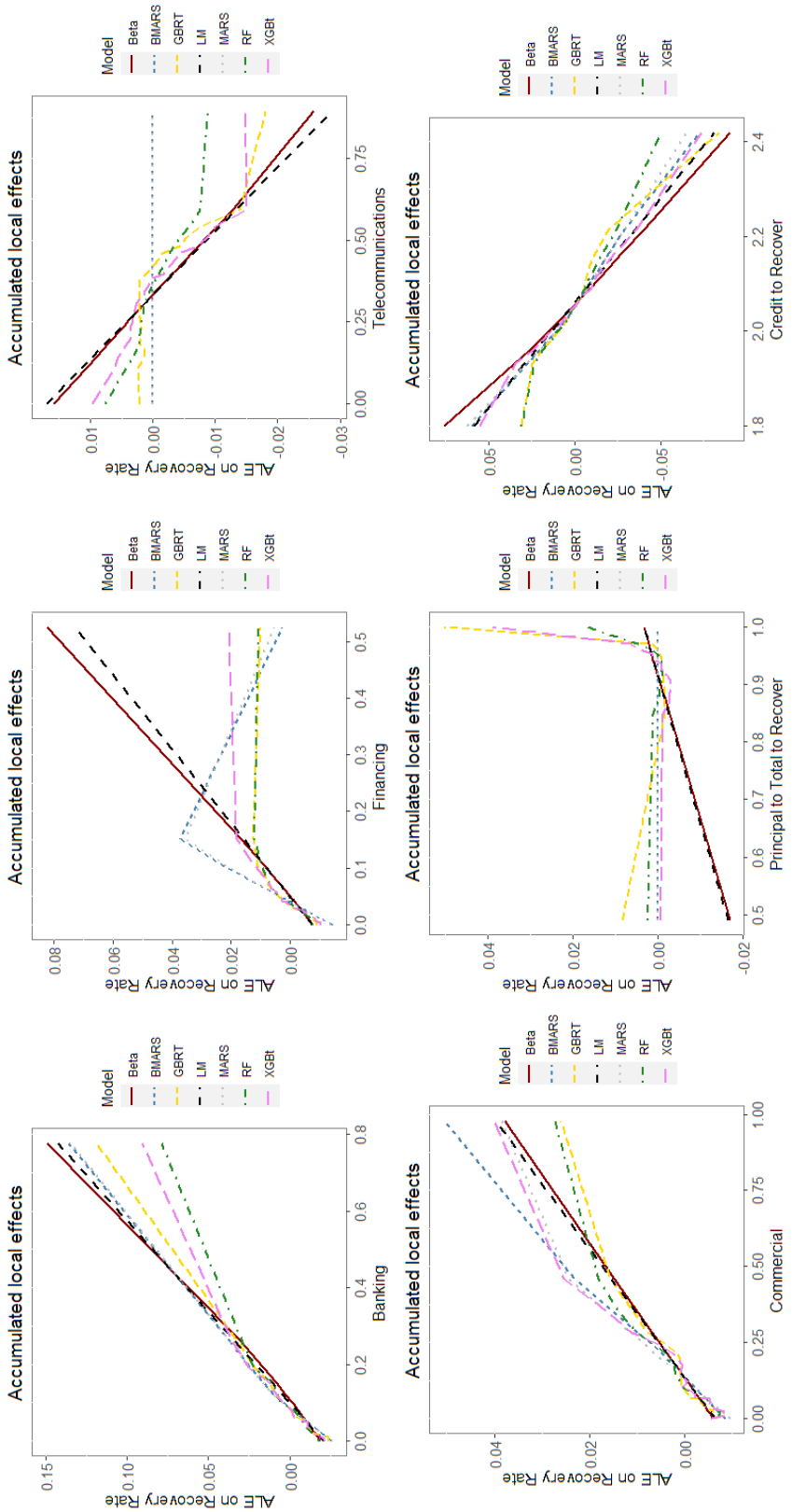


Figure 4.10: Accumulate Local Effect Plots.

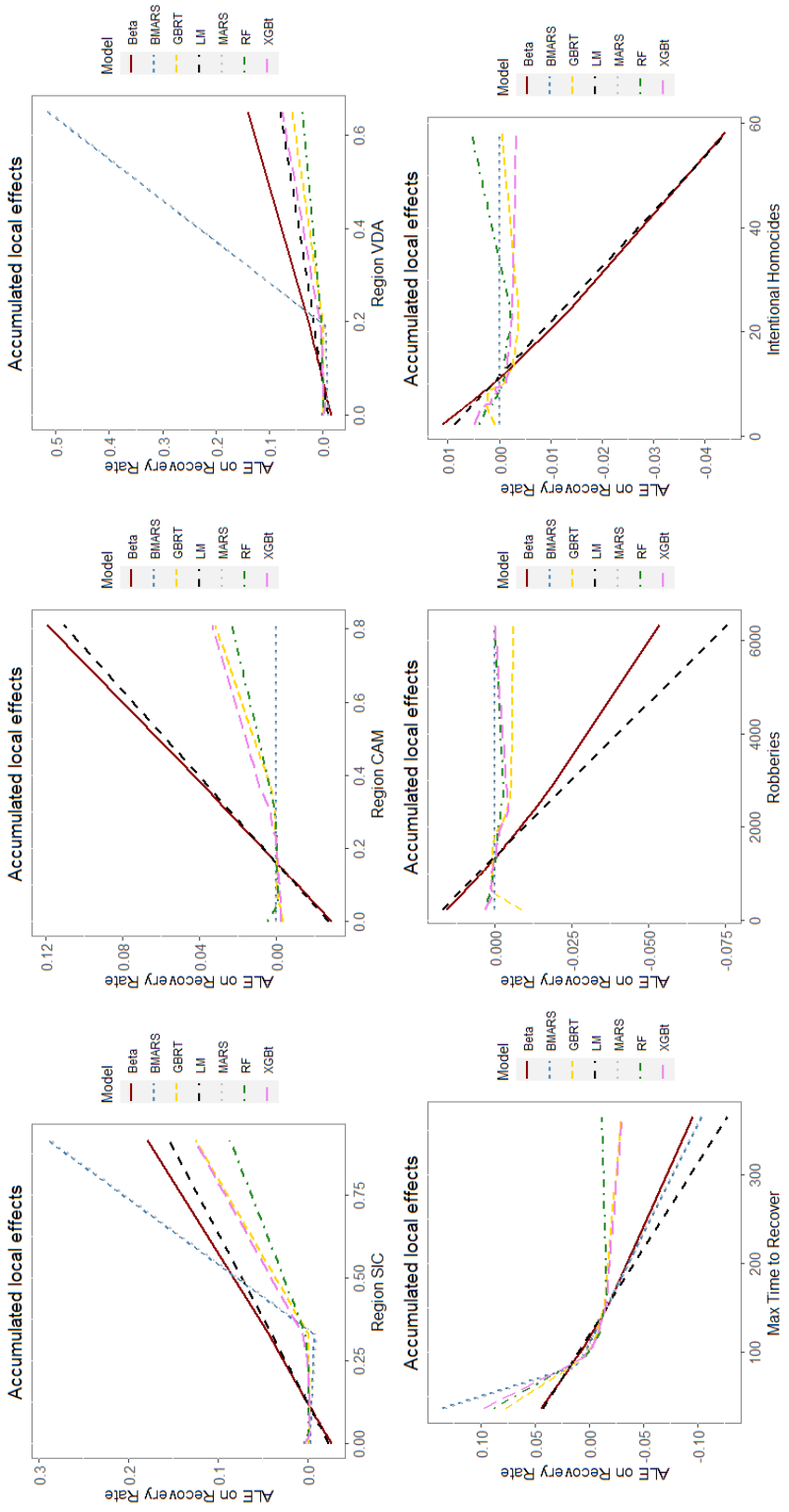


Figure 4.11: Accumulate Local Effect Plots.

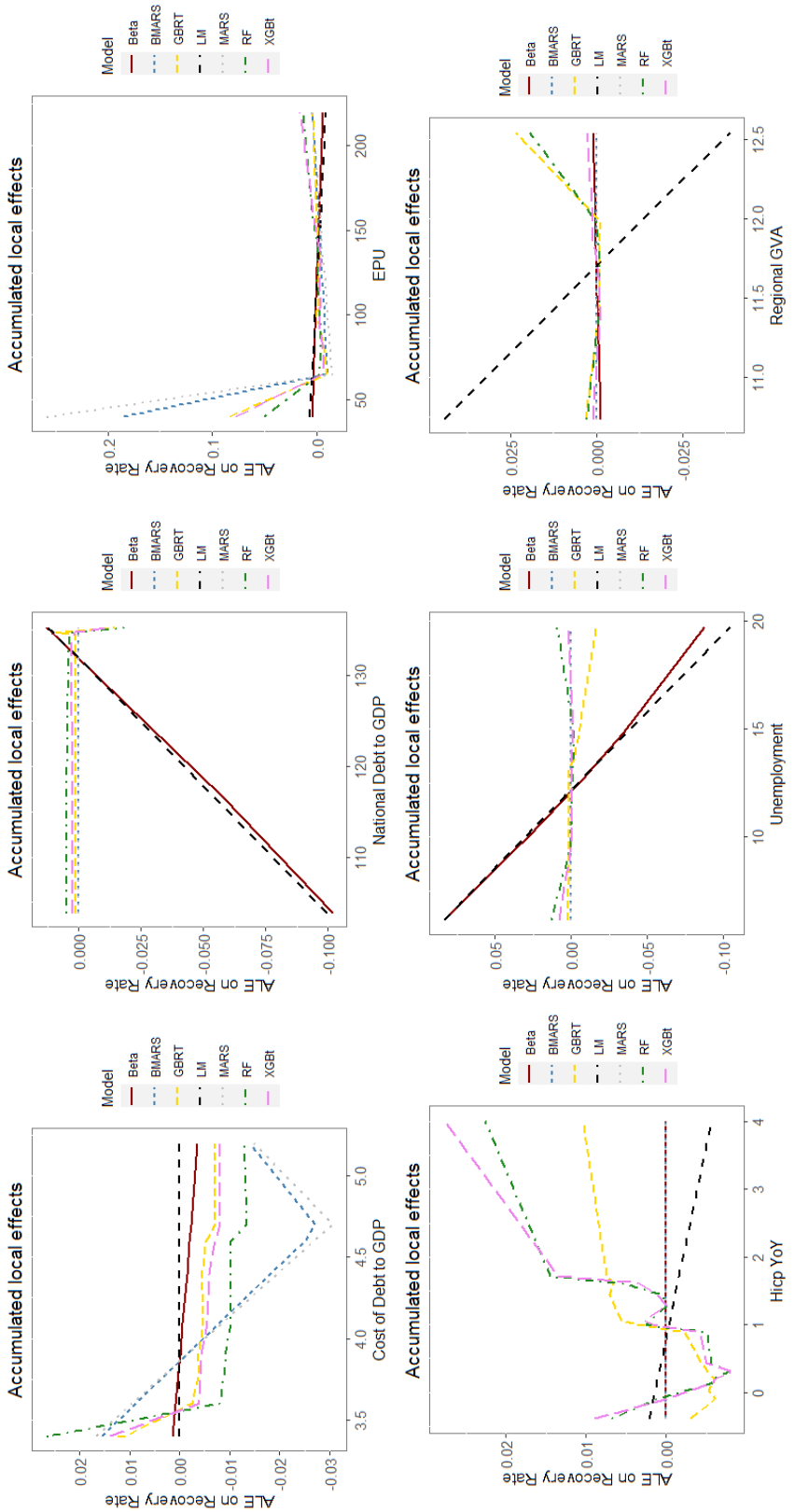


Figure 4.12: Accumulate Local Effect Plots.

APPENDIX

A4.1 Pseudo R code for hyperparameters tuning and forecast combination methods

In this section, we provide the details of the estimation of the methods under analysis. Let `mydf_training` and `mydf_fc` be the training set and the data set on which the forecasts combination are estimated, respectively. Let `y` be the target vector, `X` the matrix of predictive variables, `F` be the matrix of available forecasts, and `FE` the corresponding forecast errors. The following pseudo-code replicates the $COSE^{+}$ optimal and robust combination scheme introduced by Roccazzella, Gambetti, and Vrins (2022b) and the nonlinear combination schemes employed in this study. All models were trained befitted from parallel computing to improve computational time and final predictions have been capped in the $[0, 1]$ interval. Due to computational limitations, we fixed hyperparameters from the first iteration.

A4.1.1 Training of RF, GBRT, MARS and B-MARS models

```
######### Stochastic Gradient Boosting #########

gbm_grid <- expand.grid(.interaction.depth = seq(5, 80, by = 5),
.n.trees = c(10, 50, 100, 200),
.shrinkage = c(0.001, 0.01, 0.1, 0.2, 0.3, 0.5)
.n.minobsinnode=10)

gbm_fit <- train(as.formula("y~."),
data = mydf_training,
method = "gbm",
tuneGrid = gbm_grid,
distribution = 'gaussian',
trControl = trainControl(method = "cv",
number = 5))

######### XGB Trees #########

ALPHA   <- seq(0, 1, by = 0.05)   L1 regularization
LAMBDA  <- seq(0, 1, by = 0.05)   L2 regularization
ETA     <- seq(0, 0.5, by = 0.05) Learning Rate
GAMMA   <- seq(0, 0.5, by = 0.05) Minimum Loss for a New Leaf
SUBsamp <- seq(5, 1, by = 0.1)    Row Sampling

# Iterate over the grids of hyperparameters for ALPHA, LAMBDA, ETA, GAMMA
# and SUBsamp and take the specification that minimizes the RMSE

xgbTrees.cv <- xgb.cv(seed = 42,
```

```
max_depth = 200,
eta = ETA[i],
lambda = LAMBDA[j],
alpha = ALPHA[z],
gamma = GAMMA[t],
subsamp = SUBsamp[h],
nthread = 6,
nrounds = 50,
objective = "reg:squarederror",
data = xgb_train,
nfold = 10,
showsd = TRUE,
stratified = TRUE,
metrics = "rmse",
verbose = TRUE,
print_every_n = 1L,
early_stopping_rounds = TRUE,
maximize = FALSE )

# Best tune for individual loans
xgbTrees <- xgboost(data = xgb_train, objective = "reg:squarederror",
max_depth = 200,
eta = 0.1,
lambda = 1,
alpha = 0.62,
gamma = 0.01,
min_child_weight = 10,
subsample = 0.9,
nthread = 6, nrounds = xgbTrees.cv$best_iteration)

######### Random Forest #########
rf_grid <- expand.grid(.splitrule = 'variance',
.mtry = seq(5, 40, by = 5),
.min.node.size = c(10))

rf_fit <- train(as.formula("y~."),
data = mydf_training,
method = 'ranger',
num.trees = 5000,
tuneGrid= rf_grid,
importance = 'impurity',
trControl = trainControl(method = "cv",
number = 5))

######### MARS #########
mars_grid <- expand.grid(degree = 1:3,
nprune = 1:100)
```

```
MARS_fit <- train(as.formula("y~."),
data=mydf_training,
method = "earth",
metric = "RMSE",
tuneGrid = mars_grid,
trControl = trainControl(method = "Cv",
number = 5))

######### Bagged MARS #########

BMARS_fit <- train(as.formula("y~."),
data=mydf_training,
method = "bagEarth",
B = 30,
metric = "RMSE",
tuneGrid = mars_grid,
trControl = trainControl(method = "Cv",
number = 5))
```

A4.1.2 Minimum variance combination of forecasts with non negative weights

Using the `R` package `quadprog`[17], the function `Weights_P` estimates the combining weights given the covariance matrix `Qmat`.

```
Weights_P <- function(Qmat){
m <- ncol(Qmat)
Eq <- ones <- rep(1,m)
Ineq <- diag(1, m)
b <- c(1, rep(0,m))
dvec =  as.matrix(rep(0, m), nrow = m)
Amat = t(rbind(Eq,Ineq))
bvec = as.matrix(b, nrow =m)
meq = 1
Weights <- solve.QP(Qmat, dvec, Amat, bvec, meq)
return(Weights$solution)}
```

CO+ Constrained Optimization.

```
S <- cov(FE)
Weights_S <- Weights_P(S)
```

COSE+ Constrained Optimization with Shrinkage toward diagonal matrix[18]

```
S_shrink_D <- covEstimation(as.matrix(FE), control = list(type ='diag'))
Weights_shrink_D  <- Weights_P(_shrink_D)
```

[17] `https://cran.r-project.org/package=quadprog`.
[18] Required packages: `Ardia2017` (Ardia, Boudt, and Gagnon-Fleury, 2017)

A4.1.3 Non linear combination schemes

Using the `R` library `caret` (Kuhn, 2008) and its dependencies, we can estimate NN_{FC}. The combined forecast can be simply obtained with `predict`.

```
nn_grid <- expand.grid(decay = seq(0, 0.3, by = 0.01),
size = c(10, 20, 30, 50, 100), bag = TRUE)

nn_fc <- train(as.formula("y~."),
data=mydf_fc,
method = "avNNet",
tuneGrid = nn_grid,
maxit = 5000,
MaxNWts = 800,
repeats = 20,
trControl = trainControl(method = "Cv",
number = 5))
```

A4.1.4 Distribution of recovery rates on pools

Let N_i and R_i be the outstanding amount (i.e., amount to recover) and recovery rate associated with credit i, respectively. The average recovery rate on a pool of defaulted credits $1, \ldots, n$ is

$$\langle R_n \rangle = \frac{\sum_{i=1}^n N_i R_i}{\sum_{i=1}^n N_i} = \sum_{i=1}^n w_i R_i \ , \quad w_i = N_i \Big/ \sum_{j=1}^n N_j \ .$$

We consider a one-factor latent variable model, i.e., the R_i's are modeled as random variables taking values in $[0,1]$ and being mutually independent conditional upon a single systematic risk factor, Z. It is well known from the strong law of large numbers that, conditional upon Z an under technical assumptions[19], the variable $\langle R_n \rangle$ converges to its conditional expectation $\langle R_n(Z) \rangle := \mathbb{E}[\langle R_n \rangle | Z]$ as $n \to \infty$ (see Gordy, 2003, or details). This result essentially says that as the exposure share w_i of each asset in the portfolio goes to zero, idiosyncratic risk in portfolio loss is diversified away perfectly, so that the average recovery rate becomes a deterministic function of the systematic factor, Z. We conclude from this result that the distribution of $\langle R_n \rangle$ can be approximated reasonably well by that of $\langle R_n(Z) \rangle$, for n large enough.

For the sake of illustration, consider an equally-weighted homogeneous pool, with $w_i = 1/n$ and where the recovery rates R_i are modeled as Bernoulli random variables with parameter $\pi = \mathbb{P}(R_i = 1)$. In this illustration, we consider the one-factor Gaussian copula dependence scheme, which is the corner stone of the loss model adopted in the Basel regulation on capital requirements[20] :

$$R_i = 1_{\{X_i \leq \Phi^{-1}(\pi)\}} \quad \text{where} \quad X_i = \sqrt{\rho} Z + \sqrt{1-\rho} \tilde{X}_i$$

and $Z, \tilde{X}_1, \ldots, \tilde{X}_n$ are i.i.d. standard Normal variables with cumulative distribution function Φ. Then, from the linearity of the conditional expectation operator,

$$\langle R_n(Z) \rangle = \mathbb{E}[\langle R_n \rangle | Z] = \sum_{i=1}^n w_i \mathbb{E}\left[1_{\{X_i \leq \Phi^{-1}(\pi)\}} | Z\right]$$
$$= \frac{1}{n} \sum_{i=1}^n \mathbb{P}(X_i \leq \Phi^{-1}(\pi) | Z) = \Phi\left(\frac{\Phi^{-1}(\pi) - \sqrt{\rho} Z}{\sqrt{1-\rho}} \right) \ .$$

This leads to the following approximation for the distribution of $\langle R_n \rangle$ for large n:

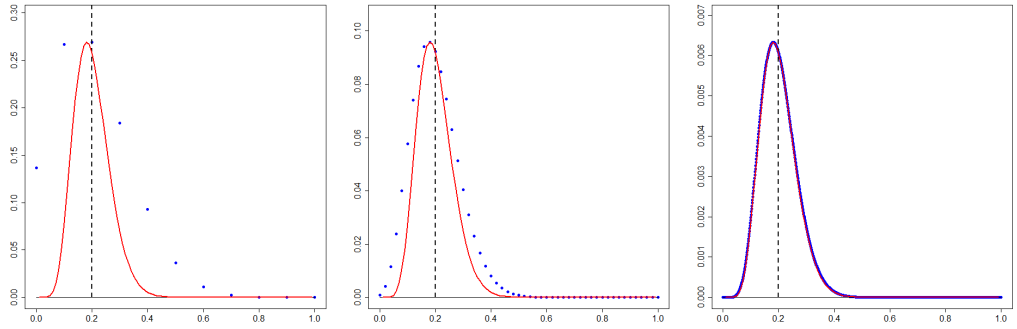
$$\mathbb{P}(\langle R_n \rangle \leq x) \approx \mathbb{P}(\langle R_n(Z) \rangle \leq x)$$

where, using $\Phi(x) = 1 - \Phi(-x)$),

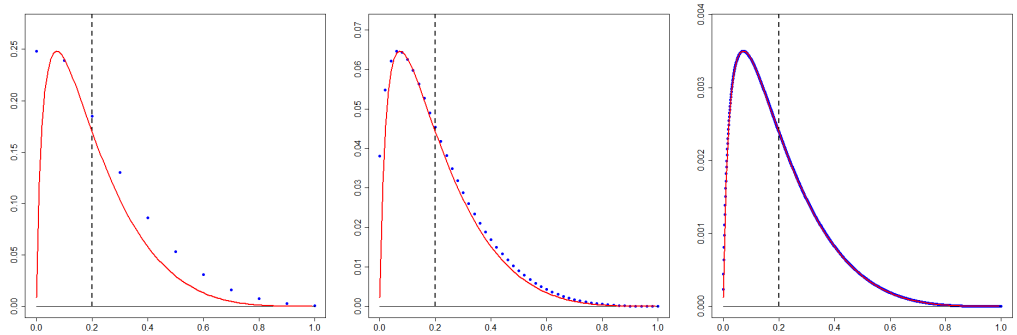
$$\begin{aligned} \mathbb{P}(\langle R_n(Z) \rangle \leq x) &= \mathbb{P}\left(\Phi\left(\frac{\Phi^{-1}(\pi) - \sqrt{\rho} Z}{\sqrt{1-\rho}} \right) \leq x \right) \\ &= \mathbb{P}\left(Z \geq \frac{\Phi^{-1}(\pi) - \sqrt{1-\rho} \Phi^{-1}(x)}{\sqrt{\rho}} \right) \\ &= \Phi\left(\frac{\sqrt{1-\rho} \Phi^{-1}(x) - \Phi^{-1}(\pi)}{\sqrt{\rho}} \right) \ . \end{aligned}$$

[19] including that $\sum_{i=1}^n N_i \to \infty$ and $\max_{i \in \{1,2,\ldots,n\}} w_i = O(n^{-(1/2+\zeta)})$ for some $\zeta > 0$ as $n \to \infty$.

[20] In Basel regulation, the goal is to model the default events, i.e., the default indicators. However, in this setup where the recovery rates are assumed to be binary, a similar approach can be used to couple the recovery rates. Indeed, the recovery rates are known to be negatively correlated with the default rate, hence, exhibit a positive dependence.



(a) $\rho = 5\%$



(b) $\rho = 25\%$

Figure A4.1: Exact (blue dots) and asymptotic (red solid line) distributions of the average recovery rate on a equally-weighted homogeneous pool ($w_i = 1/n$, $\pi = 0.2$) of size n in a single factor Gaussian model for $n = 10, 50, 1000$ (left to right). The asymptotic distribution provides a good approximation even for relatively small n (a few tenths). Note that the density curve (in red) is rescaled so that its maximum matches with the maximum of the probability mass function of $\langle R_n \rangle$.

Noting φ the density of the standard Normal random variable, the asymptotic approximation for the averaged recovery rate density is

$$\hat{f}_{\langle R_n \rangle}(x) := \frac{1}{\varphi(\Phi^{-1}(x))} \sqrt{\frac{1-\rho}{\rho}} \varphi\left(\frac{\sqrt{1-\rho}\Phi^{-1}(x) - \Phi^{-1}(\pi)}{\sqrt{\rho}} \right) \tag{A4.1}$$

Figure A4.1 illustrates the true and asymptotic distributions of the averaged recovery rate on an equally-weighted homogeneous pool in a single factor model and Gaussian coupling scheme with $\pi = \mathbb{E}[R_i] = 0.2$ for various values of ρ and n. The exact distribution of $\langle R_n \rangle$ is discrete, and the corresponding probability mass function is obtained by noting that conditional upon Z, the average recovery rate follows a Binomial law. The unconditional law of $\langle R_n \rangle$ is found by integrating Z out, which is performed numerically, using Gaussian quadrature. The asymptotic density in eq. (A4.1) offers quickly a good approximation of the exact distribution associated with the recovery rate model (in practice, as from n is equal to a few tenths).

Chapter 5

Should we care about ECB inflation expectations?

with Bertrand Candelon

Abstract

We evaluate time-series and institutional inflation projections in the euro area using optimal forecast combinations. Combination weights reveal that the ECB is the most informative forecaster on average, but it does not encompass its competitors and its weight varies over time. Macro-financial conditions explain this variability. The greater the uncertainty surrounding inflation, the difference between the current and 2% inflation target, and the more loosening the monetary policy, the less informative the ECB is. The recent decline of the ECB weight signals a potential loss of informativeness, damaging the ECB's ability to anchor expectations and questioning whether pursuing financial stability is compatible with inflation targeting.

Introduction

Forecasting inflation is essential for policymakers (central banks or treasury offices) to fix their policies but also for investors interested in expected real return, labor market actors during the wage negotiation processes, or households and firms, desiring to evaluate their purchasing power, their saving capacities or planning their inventory. Therefore, it is not surprising that many inflation forecasts are available. On the one hand, we can find multiple forecasts building upon time-series models (Stock and Watson, 1999). These models constitute the reduced forms of well-known relationships (Phillip's curve, the slope of the yield curve,...) and provide the best prediction at time t given the available information. On the other hand, institutional forecasts are also widely available and these rely on both quantitative forecasts and expert judgment. For example, the Survey of Professional Forecasters (SPF) and other international institutions such as the European Central Bank (ECB), the International Monetary Fund (IMF), the European Commission (EC), and the Organization for the Economic Cooperation and Development (OECD) are proposing regular projection exercises to forecast inflation future dynamics.

An extensive body of research compares time-series-based and institutional forecasts. For example, Stock and Watson (1999) highlight that the SPF inflation forecasts for the U.S. are dominated by a simple reduced form of the structural Phillips curve forecast. However their conclusion has been widely criticized, showing that this result is not robust to the sample data and the evaluations techniques used. For example, Ang, Bekaert, and Wei (2007) propose a combination forecast exercise including both the SPF and the time series indicators. In this perspective, the forecast combination is not a goal

per se, but the combining weights are estimated to test for forecast encompassing, i.e., investigating whether a forecast combination incorporates more information than the individual forecasts (Clements and Hendry, 1998). Similarly to finance's portfolio theory, the weights reflect forecasts' marginal contribution to the reduction of the forecaster's loss function (Roccazzella, Gambetti, and Vrins, 2022b) and, therefore, when looking at forecast encompassing and forecast combination from a joint perspective, weights can be naturally interpreted as the marginal contribution of each forecaster in term of independent information. In this context, Ang, Bekaert, and Wei (2007) show that the SPF is more informative on US inflation, i.e., the SPF either encompasses or displays significantly greater weight than the competing time-series forecasts when included in a combination. In the euro area (EA), Diebold and Shin (2019) bring evidence that optimal combining weights for the EA inflation projections produced by individual forecasters within the SPF vary over time. This finding reveals that also the informational value embedded in such forecasts is equally nonconstant over time. It remains still unknown whether this result holds for a broader set of forecasters, including institutions such as central banks or the IMF.

We propose to shed new light on the informativeness of inflation forecasts in the euro area by constructing a novel optimal combination of forecasts between these existing indicators. We constrain the weights to lay in the unit simplex in the spirit of Conflitti, Mol, and Giannone (2015) and design a forecast encompassing test to assess whether the weights are statistically significant.

This chapter extends this literature in three ways. First, together with the time series indicator and the surveys, it includes inflation forecasts published by international institutions. We provide additional insights on the relative importance of well-known international organizations *vis-à-vis* of the traditional time series and surveys for inflation forecasting. Second, this chapter adapts statistical testing under constrained parameter space to forecast encompassing. By explicitly considering that weights are constrained in the unit simplex, the proposed test displays adequate size properties and it is more powerful than a competitive test that ignores such constraints. Furthermore, the testing procedure can be extended to frameworks in which the number of competing forecasts outclasses the number of available realizations of the target variable. Third, while understanding the determinants of the weights' dynamics is interesting for macroeconomic analysis *per se*, studying in what economic states the ECB inflation forecast is more informative also helps policy analysts interested in evaluating the ECB's intent at monitoring inflation. The role of the ECB is very peculiar: it provides regular inflation forecasts but it also enhances the monetary policy to maintain price stability. Therefore, the inflation forecasts produced by the central bank should naturally be the most informative as they may contain private information on the future monetary policy (Svensson, 2005; Faust and Wright, 2013). Despite the primary objective of the ECB mandate is monitoring inflation dynamics, it also aims at preserving financial stability. A declining information content of the ECB inflation forecasts might signal that the objective of financial stability can challenge the inflation targeting mandate. We, therefore, investigate whether macro-financial conditions and monetary policy actions affect the informativeness of the ECB forecasts.

We find that the ECB is the most informative forecaster on average. As monetary

policy led by the ECB relies on their internal projection, the ECB forecasts must have an intrinsic value. However, no individual forecasts of inflation (ECB included) encompass the competing forecasts in the euro area. We also bring evidence that the optimal weights greatly vary over time. This indicates that the informational value embedded in the forecast of interest is equally nonconstant over time. This creates an ideal laboratory to study whether uncertainty, monetary policy, or macro-financial conditions in the euro area explain its variability. Finally, the ECB declining weight and the relation to its determinants raise a warning flag. This can potentially damage the ECB's leading role in anchoring inflation expectations and can signal a possible incompatibility between preserving financial stability and the inflation targeting goal in a high inflation environment.

The remainder of the chapter is organized as follows. We introduce the optimal combination of forecasts and the forecast encompassing test in Section 5.1. We move to the analysis of inflation forecasts in the euro area in Section 5.2 and we study the dynamics of the weights in Section 5.3. We conclude in Section 5.4.

5.1 Optimal combination and forecast encompassing

In the following sections, we introduce the ex-post optimal combination of forecasts and the testing procedure to evaluate their significance in a constrained parameter space.

5.1.1 Estimating the optimal weights

Assuming unbiased forecasts and under a quadratic loss function, an analyst may seek to minimize the variance of forecast errors.[1] Let $\boldsymbol{y}$ be the column vector containing n observations of the target variable.[2] Let $\boldsymbol{f}_i$ be the vector of unbiased forecast of $\boldsymbol{y}$ produced by forecast i and define $\boldsymbol{F}$ the $n \times d$ matrix containing the d forecasts of $\boldsymbol{y}$. Similarly, we can define $\boldsymbol{V}$ as the $n \times d$ matrix of forecast errors and $\boldsymbol{S}$ to be an estimator of the covariance matrix of prediction errors. We denote the population covariance matrix of prediction errors with $\boldsymbol{\Sigma}$. We aim at estimating the linear combination of forecasts that minimizes the variance of the aggregate prediction error constraining the weights to be in the unit simplex. Formally, denoting the unit simplex with d vertices with $\Delta^{d-1} = \{\beta_i \geq 0, \text{ with } i = 1, ..., d \text{ and } \sum_{i=1}^{d} \beta_i = 1\}$, the optimal nonnegative weights solve

$$\widehat{\boldsymbol{\beta}} := \underset{\boldsymbol{\beta} \in \Delta^{d-1}}{\arg\min}\, \boldsymbol{\beta}^T \boldsymbol{S} \boldsymbol{\beta} \,, \tag{5.1}$$

The nonnegativity constraint offers three advantages: a) it rules out extreme solutions that result from estimation error when dealing with the minimum-variance objective function in small samples (Jagannathan and Ma, 2003); b) it identifies noise forecasts, i.e., predictions that individually do not help to predict the target, but whose inclusion in the combination of forecasts is only justified by the objective of minimizing the *in-sample* variance of prediction errors[3]; and c) in this context, the estimator of the weights turns into a special case of the Lasso (Tibshirani, 1996), which permits us to extend the test

[1] The inclusion of an intercept allows biased forecasts to be evaluated (Timmermann, 2006).

[2] In this chapter, we denote matrices with capital bold letters, vectors with lowercase bold letters, and scalars with lowercase letters.

[3] It is well known that the inclusion of such forecasts can negatively impact the stability of the weights (Bunn, 1981).

for forecast encompassing to high dimensional problems by encouraging sparse portfolios (i.e., portfolios with only a few active positions) (Brodie, Daubechies, De Mol, Giannone, and Loris, 2009). Indeed, Fan, Zhang, and Yu (2012) employed this correspondence to estimate approximate weights for the minimum variance portfolio under gross-exposure constraints when the number of securities under consideration is (potentially) larger than the sample size.

5.1.2 Testing for forecast encompassing

We next turn to design a test for multiple forecasts encompassing. Granger and Ramanathan (1984) show that (5.1) is equivalent to a linear regression problem, where the weights are estimated using

$$\widehat{\boldsymbol{\beta}} = \underset{\beta \in \Delta^{d-1}}{\arg\min} \frac{1}{n} \left[\boldsymbol{\epsilon}^T \boldsymbol{\epsilon} \right] \ , \tag{5.2}$$

with $\boldsymbol{\epsilon} := \boldsymbol{y} - \boldsymbol{F}\boldsymbol{\beta}$ and under the constraint $\boldsymbol{\beta}^T \mathbf{1} = 1$.

Noticing that weights are constrained to sum to one, we can rewrite (5.2) as a linear regression problem where the forecast error corresponding to the i^{th} forecast becomes the dependent variable and the remaining $\boldsymbol{V}_{\cdot,z}$ with $z = 1, 2, ..., d-1$ become the explanatory variables (Fan, Zhang, and Yu, 2012), i.e.,

$$V_d = \sum_{i=1}^{d-1} \beta_i \boldsymbol{V}_{\cdot,i} + \boldsymbol{u}, \quad \text{where} \quad \mathbb{E}[\boldsymbol{u}|\boldsymbol{V}_1, \boldsymbol{V}_2, ..., \boldsymbol{V}_{d-1}] = 0, \quad \mathbb{E}[\boldsymbol{u}\boldsymbol{u}^T] < \infty \ , \tag{5.3}$$

or alternatively,

$$V_i = \boldsymbol{V}_{\setminus i} \boldsymbol{\beta}_{\setminus i} + \boldsymbol{u} \ . \tag{5.4}$$

where $\boldsymbol{V}_{\setminus i}$ is the $n \times d-1$ matrix containing the remaining forecast errors and $\boldsymbol{\beta}_{\setminus d}$ is the $d-1$ vector of weights contains all but the weight corresponding to the forecast i^{th}. The weight associated with the i^{th} forecast can be trivially retrieved: $\beta_i = 1 - \boldsymbol{\iota}^T \boldsymbol{\beta}_{\setminus i}$.

In the spirit of Harvey and Newbold (2000), we test whether the marginal contribution of forecast i to the aggregate ex-post optimal forecast is statistically significant. However, this test must explicitly consider the constraints on the weights, i.e., we state the following hypothesis

$$H_0 : \boldsymbol{\theta} = 0 \quad vs \quad H_1 : \boldsymbol{\theta} > 0, \quad \boldsymbol{\theta} \in \mathbb{R}^{+(r)} \ , \tag{5.5}$$

where we consider the following partition $\boldsymbol{\beta} = (\boldsymbol{\theta}, \boldsymbol{\gamma})$, with $\boldsymbol{\theta}$ being the $r-$dimensional parameter of interest and $\boldsymbol{\gamma}$ denoting the nuisance parameters (with dimension $d-r$). This corresponds to test whether the population ex-post optimal weights lay on the boundary set (under H_0) or in the cone[4] generated by the nonnegativity and the sum to one constraints (under H_1). In this chapter, we assume that each realization of $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon)$[5] and we denote the sample negative log-likelihood function as

$$\ell(\boldsymbol{\beta}) = -\frac{1}{n} \sum_{i=1}^{n} \log \mathcal{L}_i(\boldsymbol{\beta}) = \frac{n}{2} log(\pi \sigma_\epsilon) + \frac{1}{2} \frac{[\boldsymbol{y} - \boldsymbol{F}\boldsymbol{\beta}]^T [\boldsymbol{y} - \boldsymbol{F}\boldsymbol{\beta}]}{\sigma_\epsilon} \ , \tag{5.6}$$

[4] A subset C of $\mathbb{R}^d$ is called a cone if $\forall x \in C$, $tx \ \in C$ for every positive t.

[5] We discuss the relevance of the normality assumption for our empirical application in A5.4.

with the corresponding likelihood ratio statistics defined by

$$T_{LR} = 2n \left(\ell(\widehat{\boldsymbol{\beta}}_{H_0}) - \ell(\widehat{\boldsymbol{\beta}}_{H_1}) \right) , \tag{5.7}$$

or the asymptotically equivalent score statistics[6]

$$T_S = \left[\nabla \ell(\widehat{\theta}_{H_0}) - \nabla \ell(\widehat{\theta}_{H_1}) \right]^T \widetilde{\mathscr{F}}_{\theta|\gamma}^{-1} \left[\nabla \ell(\widehat{\theta}_{H_0}) - \nabla \ell(\widehat{\theta}_{H_1}) \right] . \tag{5.8}$$

Under these assumptions, Kudô (1963) proved that, under the null hypothesis, the LR and score statistics associated to (5.5) are distributed as a mixture of chi-squared distributions. Determining the mixing weights is essential to compute the cumulative tail probability of the test statistics for a given significance level a. When the constrained space is the nonnegative orthant, Silvapulle and Sen (2004) propose a simple quadratic programming problem to estimate the vector of mixing weights ω.[7] When θ is univariate ($r = 1$), the weights are known in analytical form and the chi-bar-squared statistics reduce to

$$\text{Prob}\left[\overline{\chi}^2(\widetilde{\mathscr{F}}_{\theta|\gamma}, \mathbb{R}^+) > c_a \right] = .5 \text{ Prob}\left[\chi^2(0) > c_a \right] + .5 \text{ Prob}\left[\chi^2(1) > c_a \right] , \tag{5.9}$$

where $\chi^2(df)$ denotes a chi-squared distribution with df degrees of freedom. Finally, given a significance level a and its corresponding critical value c_a, we compute cumulative tail probability as:

$$\text{Prob}\left[\overline{\chi}^2(\widetilde{\mathscr{F}}_{\theta|\gamma}, \mathbb{R}^{+r}) > c_a \right] = \sum_{i=0}^{r} \omega_i \text{Prob}\left[\chi^2(i) > c_a \right] ,$$

and we reject H_0 if

$$\text{Prob}\left[T_S > c \right] > a \quad \text{or} \quad \text{Prob}\left[T_{LR} > c \right] > a \, .$$

Large data sets, heterogeneous modeling techniques, and, different data pre-processing procedures originate a wide array of forecasts, whose size can potentially outclass the number of available realizations of the target variable. Therefore, a "modern" test for forecast encompassing should also deal with potential high dimensional frameworks. This test can accommodate these features using the decorrelated score procedure of Ning and Liu (2017) as described by Yu, Gupta, and Kolar (2019).

In the remainder of the chapter, to ease presentation and simplify notation, we focus on the likelihood framework in a low dimensional setting and on the uni-variate parameter of interest θ ($r = 1$) corresponding to the j^{th} forecast. We refer to Appendix A5.1 and A5.2 for the extension of the test in the high dimensional case and Appendix A5.7 for the power analysis and the study of the size properties of the test.

[6] We let $\nabla \ell(\boldsymbol{\beta}) = \nabla \ell(\theta, \boldsymbol{\gamma})$ be the gradient of the negative log likelihood, $\nabla^2 \ell(\boldsymbol{\beta})$ be the sample Hessian matrix and $\mathscr{F}(\beta) = \mathbb{E}[\nabla^2 \ell(\boldsymbol{\beta})]$ be the population Fisher information matrix, while we denote with $\nabla_\theta \ell(\boldsymbol{\beta})$, $\nabla_\gamma \ell(\boldsymbol{\beta})$, $\mathscr{F}(\beta)_{\theta\theta}$, $\mathscr{F}(\beta)_{\theta\gamma}$, $\mathscr{F}(\beta)_{\gamma\theta}$, $\mathscr{F}(\beta)_{\gamma\gamma}$ the correspondent partitions. We also denote with $\widehat{\boldsymbol{\beta}} = (\widehat{\theta}, \widehat{\gamma})$ the estimated ex-post optimal weights.

[7] See proposition 3.6.1 of Silvapulle and Sen (2004).

5.2 Analysis of inflation forecasts in the euro area

We next turn to our empirical study of the dynamics of yearly inflation forecasts in the euro area for the years ranging from 2009 to 2021. We have a total of ten forecasts: five are produced by institutions or professional forecasters, one is based on a random walk (RW), one relies on autoregressive integrated moving average ($ARIMA$) models, and three bi-variate vector-autoregressive (VAR) models.

To ease the presentation, we classify forecasts into institutional and model-based. We follow Patton and Timmermann (2010) and use the December-on-December change of the realized value of HICP to measure yearly inflation. The HICP growth rate is the official measure of inflation in the euro area and the target of the considered institutional forecasts of inflation.[8] For each quarter, we consider inflation forecasts for the same calendar year. Indeed, compared to standard forecasting exercises, in the case of institutional forecasts, the horizon is not fixed, e.g., 1 year ahead. Forecasters predict the HICP growth rate of a certain calendar year, e.g., the yearly inflation rate in 2018. We avoid longer-term forecasts (1 and 2 calendar years ahead) because institutional forecasters fail to systematically update their predictions beyond one calendar year ahead (Andrade and Le Bihan, 2013; Easaw and Golinelli, 2021).

Financial instruments such as inflation-linked swaps (ILS) and inflation-linked bonds (ILB), such as the Italian BTPi indexed Treasury bonds can also be used as a market-based source of inflation forecasts. In the euro area, the Harmonised Index of Consumer Prices excluding tobacco is the relevant index to which both ILS and ILB are anchored. This is a shortcoming of using ILS and ILB-implied inflation forecasts, as they forecast a different measure of inflation compared to the ECB, the SPF, the EC, the IMF, and the OECD that, instead, consider the HICP index. Moreover, ILB and ILS yields embody not only the inflation expectation, but also a risk premium related to inflation uncertainty, counterparty, and (in the case of ILB) credit risk, which can vary over time. ILB and ILS inflation-based forecasts cannot be directly employed without disentangling all these components. For these reasons, ILS and ILB implied inflation forecasts cannot be employed in this study.

Institutional forecasts data are retrieved from *European Central Bank staff macroeconomic projection for the euro area*[9] and consist of forecasts produced by the ECB staff, the OECD, the SPF, the IMF, and the EC. The ECB, the EC, and the SPF release their inflation forecasts quarterly, while the IMF and OECD publish inflation projections at a minimum of twice per year in their respective economic outlooks. This offers an informative advantage to institutional forecasters that update their expectations more frequently.[10]

[8] In this work, we focus on point forecasts, and, as standard in the forecast encompassing literature, we assume that the official measure of inflation in the euro area, i.e, the HICP index, is observed without error. However, we remark that in the linear regression model, on which the optimal forecast combination is based, it is possible to accomodate for measurement error in the target variable if this is exogenous with respect to the considered forecasts. In this case, the estimator of the weights remains unbiased and consistent, but the variance of the residual ϵ is expected to increase.

[9] Available at `https://www.ecb.europa.eu/pub/projections/html/all-releases.en.html`

[10] This may explain why OECD and IMF forecasts display worse performance than the other institutional forecasters (see St. Dev. of the forecast errors in Table 5.1). This can also be reflected in a lower weight

Moving to our ARIMA, VARs, and RW forecasts, we follow the standard notation for the ARIMA model (Lütkepohl, 2007) and we denote with p, d, and q the order of the autoregressive component, the differencing order, and the moving average component. Formally:

$$\left(1-\sum_{i=1}^{p}\alpha_i L^i\right)(1-L)^d y_t = \left(1-\sum_{i=1}^{q} b_i L^i\right)\epsilon_t, \quad \epsilon_t \sim \ i.i.d.\ \mathcal{N}(0,\sigma^2)\ . \tag{5.10}$$

We consider three bi-variate VARs: a) we model the joint dynamics of the monthly year-on-year (y-o-y) HICP growth rate and the industrial production; and b), the y-o-y HICP growth rate and the y-o-y differences of unemployment rates; and c) the y-o-y HICP growth rate and the realized term spread of the euro area AAA government bonds at 3 months and 10 years maturity.[11] Specifications a) and b) reflect strategies that rely on the formulation of the generalized Phillips curve in terms of measures of the real activity or unemployment rate to forecast inflation in the spirit of Stock and Watson (1999) and Bekaert, Hoerova, and Lo Duca (2013). Specification c) implicitly assumes that forward-looking asset prices and, specifically, the term structure of nominal interest rates, contain relevant information to forecast inflation (Stock and Watson, 2003). The underlying economic rationale is that investment decisions rely on the real interest rate and, therefore, for nominal rates, a downward-sloping yield curve at longer maturities reflects expectations of a declining rate of inflation, while a steeply upward-sloping yield curve indicates expectations of a rising rate of inflation (Mishkin, 1990, 1991; Estrella and Mishkin, 1997). Data is retrieved from Eurostat. The inflation dynamics obtained from the bi-variate VAR(p) is:

$$\boldsymbol{y}_t = \boldsymbol{\alpha} + \sum_{i=1}^{p} \boldsymbol{A}_i \boldsymbol{y}_{t-i} + \boldsymbol{\epsilon}_t \sim \ i.i.d.\ \mathcal{N}(0,\boldsymbol{\Sigma})\ , \tag{5.11}$$

where $\boldsymbol{y}$ is the 2×1 vector containing inflation rates and either y-o-y industrial production growth rate or y-o-y differences of unemployment rate, $\boldsymbol{\alpha}$ is the 2×1 vector of the intercept and $\boldsymbol{A}_i$, $i=1,2,...,p$ are 2×2 matrices of coefficients.

We estimate the ARIMA and bi-variate VARs via maximum likelihood and consider a rolling window of 48 months.[12] For each of the considered windows, we control for the stationarity of y-o-y HICP growth rate and select the appropriate differencing order d by performing the augmented Dickey-Fuller test. Similarly, we select the appropriate (p,q) order of the ARIMA and p order of the VAR using the Akaike information criterion (AIC). We conduct a (pseudo) real-time forecasting exercise[13] with the RW, the ARIMA,

because of their lower publication frequency. However, this particular issue is not relevant in this study's context. The objective is to provide guidance to forecasters and identify the most valuable inflation forecast for the euro area, regardless of whether the advantage stems from higher publication frequency, access to private information on the expected path of monetary policy, or simply greater expertise and skill.

[11] We approximate y-o-y growth rates using log differences. This simplifies the aggregation process from monthly y-o-y growth rates to the yearly inflation measure by simply computing the arithmetic average.

[12] We have also considered rolling windows of 24, 60, and 72 months as well as expanding windows. We chose the window of 48 because it is the one that offers the best performance in terms of root mean square error (RMSE).

[13] We aim at replicating the real-time forecasting exercise as closely as possible, however, we do not have real-time vintage data sets at hand to control for possible revisions of the underlying time-series. For this reason, we call this application *pseudo* real-time.

and the VARs models. Specifically, we assume that at any month t we have access to data from the previous months on the variables of interest. Indeed, institutional forecasters are usually published at the end of the reference month and, therefore, when considering VARs, ARIMA, and RW, we construct forecasts at the end of the reference month with only data from the prior months.[14] Finally, to forecast the yearly inflation rate, we compound the observed monthly year-on-year inflation rates for the reference year with iterative forecasts for the remaining months. To illustrate this point, let's consider March 2020. In this case, we use the data for the HICP index before March 2020 to forecast the year-on-year inflation rate for March and the remaining nine months of the year. We combine these predictions with the observed year-on-year inflation rates in January and February and take the average to obtain the forecast for the yearly inflation rate.[15]

Our model-based forecasts can be updated monthly for the reference calendar year. However, institutional forecasters release their forecasts at different times during each quarter, which could give an advantage to institutions that publish their forecasts later in the quarter.[16] To address this potential bias, we calculate the optimal weights and associated statistics using the most recent forecasts made by each institution and model for each month within the quarter. Then, we average these values to determine the optimal weights and corresponding statistics assigned to each forecaster for the reference quarter, handling the adverse effects of the asynchronicity of the publication process.

Table 5.1 reports the summary statistics and Figure 5.1 depicts the dynamics of the yearly HICP growth rate and its forecasts. On average, both institutional and model-based forecasts are unconditionally unbiased and track the dynamics of yearly inflation rate but, looking at Figure 5.1, it is worth underlining three exceptions. First, forecasts based on the bi-variate VAR including industrial production tend to be more volatile, especially at the beginning of each year. This is particularly evident for VAR Slope which relies on the term spread of the nominal yields: its forecast is biased in the 2017-2021 period.

Second, most forecasts tend to overestimate yearly inflation from 2015 to 2017. Third, except for the VAR Slope, forecasts tend to either jointly overestimate or jointly underestimate inflation in the euro area, implying forecast errors to be cross-sectionally dependent. This is not surprising, especially for institutional forecasters as it is reasonable to assume that before publishing a new forecast, professional forecasters will compare their results with the ones published by other sources. For example, the ECB acknowledges the use of the SPF and market-based forecasts when forming their expectations. In the next section, we further discuss how we quantify and tackle the cross-sectional dependence problem.

[14] This assumption is fair given the monthly time-series we are considering. For example, the euro area HICP's first 'flash estimate' is published around the end of the reference month. Similarly, industrial production and unemployment are usually published 45 and 30 days after the end of the reference month, respectively.

[15] In this framework, as the year goes on, both institutional and model-based forecasters have more information and their forecasts should converge to the realized HICP yearly growth. Therefore, the largest forecast errors are most likely realized in the first two quarters of the year. Implicitly, the L2 loss function will penalize more these errors, therefore attenuating this potential problem. Evidence of potential autocorrelation in the residuals ϵ arising from this problem is also missing.

[16] For example, during Q2, the European Commission publishes its forecasts in May while the ECB in June

Table 5.1: Summary statistics of the forecasts and realized forecast errors.

Forecast Errors										
	Time-series models					Instititional forecasts				
	RW	ARIMA	VAR IP	VAR Un	VAR Slope	ECB	OECD	SPF	EC	IMF
Min	-1.80	-1.50	-1.39	-1.26	-3.43	-0.99	-1.31	-0.99	-1.91	-1.61
Max	1.15	1.62	1.60	1.46	2.42	1.06	1.86	1.66	1.16	1.66
Median	-0.17	-0.01	-0.03	-0.02	-0.32	0.00	-0.04	-0.01	-0.01	-0.01
Mean	-0.25	-0.03	-0.09	-0.07	-0.51	0.00	-0.04	0.00	-0.03	-0.02
St. Dev.	0.61	0.44	0.49	0.48	1.06	0.28	0.52	0.39	0.43	0.56
Forecasts										
	Time-series models					Instititional forecasts				
	RW	ARIMA	VAR IP	VAR Un	VAR Slope	ECB	OECD	SPF	EC	IMF
Min	0.29	0.2	-0.54	-0.51	-0.05	0.00	0.10	0.10	-0.10	0.10
Max	2.62	2.69	3.59	2.68	3.73	2.70	2.60	2.60	2.60	2.60
Median	1.58	1.45	1.37	1.34	1.73	1.40	1.40	1.40	1.50	1.40
Mean	1.52	1.30	1.36	1.33	1.77	1.27	1.31	1.27	1.30	1.28
St. Dev.	0.57	0.80	0.88	0.82	1.10	0.83	0.71	0.74	0.79	0.68

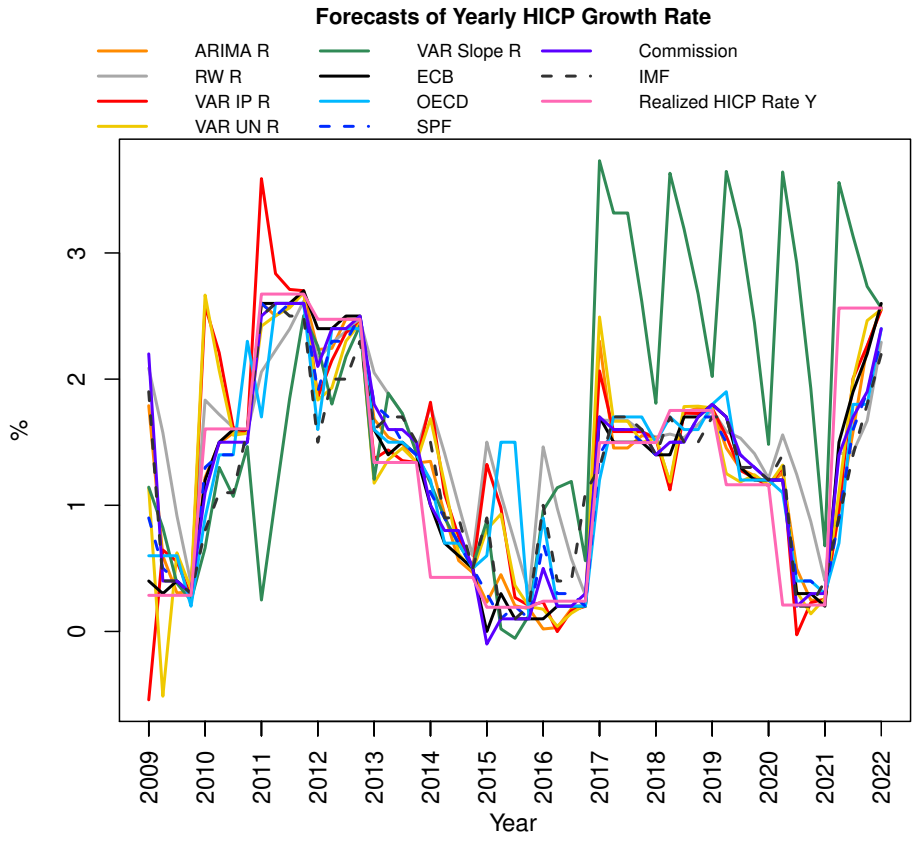


Figure 5.1: Dynamics of realized and forecasted HICP yearly growth rate.

5.2.1 Testing and handling cross-sectional dependence

Strong cross-sectional dependence of forecast error may signal potentially multi-collinear forecast errors and, consequently, imprecise estimates of the optimal weights. To quantify

the degree of cross-sectional dependence of forecast errors, we borrow the $CD-$test (Pesaran, 2021) from the panel econometrics literature, which is defined as:

$$CD = \sqrt{\frac{2T}{N(N-1)}} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \hat{\rho}_{ij}, \tag{5.12}$$

with $\hat{\rho}_{ij}$ being the sample correlation coefficient between the forecast error of model i and j. In Bailey, Holly, and Pesaran (2016), we can find the order property of the average correlation coefficient:

$$\bar{\rho}_N = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \rho_{ij} = O(N^{2\alpha-2}), \tag{5.13}$$

where $\alpha \in [0, 1]$. First, as recommended in Elhorst, Gross, and Tereanu (2020), we control that the null of weak cross-sectional dependence ($\alpha < 0.5$) cannot be rejected through the CD-test ($\hat{\bar{\rho}}$ and CD-statistic's columns in Table 5.2). We cannot reject weak cross-sectional dependence given the high value of $\hat{\bar{\rho}}$. Second, we assess the degree of cross-sectional correlation through the two-step procedure as in Bailey, Holly, and Pesaran (2016).

In Table 5.2, we show that the cross-sectional correlation of forecast errors is really strong: the estimated α is larger than .75, clearly pointing to the presence of common components driving the dynamics of forecast errors. We can reach similar conclusions by looking at the loadings of the first Principal Component shown in Table 5.2: as expected in such a case, they are evenly distributed among the different forecast errors, supporting the evidence of a persistent interdependence trend between inflation forecast errors.

5.2.2 Estimating the optimal combination weights

We estimate the forecast combination that minimizes the variance of the combined forecast error restricting the weights to be nonnegative, to sum to one and using the whole sample. We start by considering the sample maximum likelihood estimator of the covariance matrix in *Specification 1*. Indeed, as underlined in Jagannathan and Ma (2003) and Conflitti, Mol, and Giannone (2015), restricting weights to lay in the unit simplex helps reduce the risk in the estimated optimal combination of forecasts. However, being a special case of the Lasso (Fan, Zhang, and Yu, 2012), this estimator often fails to correctly select the best subset of forecasts when forecasts errors are highly correlated (Zhao and Yu, 2006). For this reason, we consider two additional estimators of the combination weights as robustness checks.

Table 5.2 suggests the presence of a strong one-factor structure characterizing the dynamics of inflation forecast errors. Following this observation, in *Specification 2*, we consider the principal orthogonal complement thresholding estimator of the covariance matrix (POET) of Fan, Liao, and Mincheva (2013). This assumes that conditional on pre-specified common factors, the residual terms are weakly correlated. Finally, we can handle the strong cross-sectional dependence by also imposing some structure on the estimator of the optimal weights. Specifically, we employ a consistent shrinkage estimator of the covariance matrix that consists of combining the sample covariance matrix (which can be easily computed and is asymptotically unbiased but prone to estimation error) with a highly structured estimator (that contains relatively little estimation error but

Table 5.2: Principal component analysis of realized forecast errors and cross-sectional dependence test.

Forecaster	PC1	PC2	PC3	PC4	PC5
ARIMA R	0.360	0.002	-0.139	0.409	-0.305
RW R	0.346	0.060	-0.012	-0.433	-0.305
VAR IP R	0.164	-0.687	0.179	0.030	0.502
VAR Un R	0.248	-0.502	0.381	0.074	-0.513
VAR Slope R	0.144	0.457	0.844	0.110	0.144
ECB	0.366	0.015	-0.010	0.084	0.073
OECD	0.338	0.065	-0.031	-0.716	0.053
SPF	0.380	0.027	-0.128	0.103	0.108
Commission	0.353	0.222	-0.241	0.301	-0.069
IMF	0.353	0.096	-0.124	0.082	0.505
Standard deviation	2.58	1.22	0.87	0.58	0.53
Proportion of Variance	66.75%	14.84%	7.56%	3.39%	2.84%
Cumulative Proportion	66.75%	81.59%	89.15%	92.54%	95.38%

$\hat{\bar{\rho}}$	CD-statistic	$\hat{\alpha}$	$\hat{\sigma}_{\hat{\alpha}}$
0.585	28.301	0.8985	0.1944

Notes. We report the estimated average sample correlation forecast errors of model i and j, $\hat{\bar{\rho}}$, the cross-sectional test statistics, $CD-$statistics, as well as $\hat{\alpha}$ and its standard error, $\hat{\sigma}_{\hat{\alpha}}$.

potentially misspecified). Being the loadings of the first principal component are roughly the same (Table 5.2), this factor can further be seen as an equally weighted portfolio of all forecasts (up to a scaling factor). Therefore, it becomes natural to shrink the sample covariance matrix towards a diagonal one (Ledoit and Wolf, 2003) that is representative of this case. We estimate the weights adopting such estimator in *Specification 3*.[17]

Table 5.3 reports the weights, the T_S score statistics defined in (5.8), and their corresponding p-values. We find that the ECB has the greatest weight across specifications, however, its forecasts do not encompass its competitors regardless of the estimation strategy. Indeed, Var IP, Var Un, and the EC forecasts are always included in the optimal combination of forecasts and their corresponding weights are statistically greater than zero. Instead, the ARIMA model, the IMF, the RW, and the VAR Slope perform poorly, being either excluded or only marginally (and insignificantly) contributing to the minimization of the aggregate prediction error.

However, Table 5.3 implicitly assumes that the performance remains constant across the whole sample. We challenge this by re-estimate the forecast combination considering a rolling window of 12 quarters.[18] Figure 5.2 displays the dynamics of the average weights

[17] Following Ledoit and Wolf (2003), the weight between the sample and the target matrix is computed by minimizing the Frobenius norm of the difference between the shrinkage estimator and the asymptotic covariance matrix.

[18] As a robustness check, we have considered a shorter window of 10 quarters. Results are comparable

Table 5.3: Ex-post optimal weights, score statistics, and p-values.

		ARIMA	RW	IP	Un	Slope	ECB	OECD	SPF	EC	IMF
Spec. 1	β	0	0	0.140	0.032	0.032	0.651	0	0	0.145	0
	T_S	.	.	4.791	4.992	2.336	5.100	.	.	6.339	.
	$P\bar{\chi}^2$	.	.	0.014	0.013	0.063	0.012	.	.	0.006	.
Spec. 2	β	0.054	0	0.118	0.091	0	0.455	0	0.170	0.088	0.023
	T_S	2.157	.	4.357	5.450	.	3.685	.	2.985	4.646	2.275
	$P\bar{\chi}^2$	0.071	.	0.018	0.010	.	0.027	.	0.042	0.016	0.066
Spec. 3	β	0	0	0.147	0.055	0.033	0.553	0	0.093	0.119	0
	T_S	.	.	4.602	5.197	2.531	3.948	.	2.630	5.561	.
	$P\bar{\chi}^2$	.	.	0.016	0.011	0.056	0.023	.	0.052	0.009	.

Notes. The sample covers the period 2012 - 2021. Spec 1: Sample Covariance, Spec 2: POET, Spec 3: L&W shrinkage toward diagonal matrix. Score statistics are computed using the realized forecast errors.

across specifications in the period Q1 2012 - Q4 2021. The x- axis identifies the reference period. For example, the interval 2015-2016 identifies the weights corresponding to the yearly inflation in 2015.

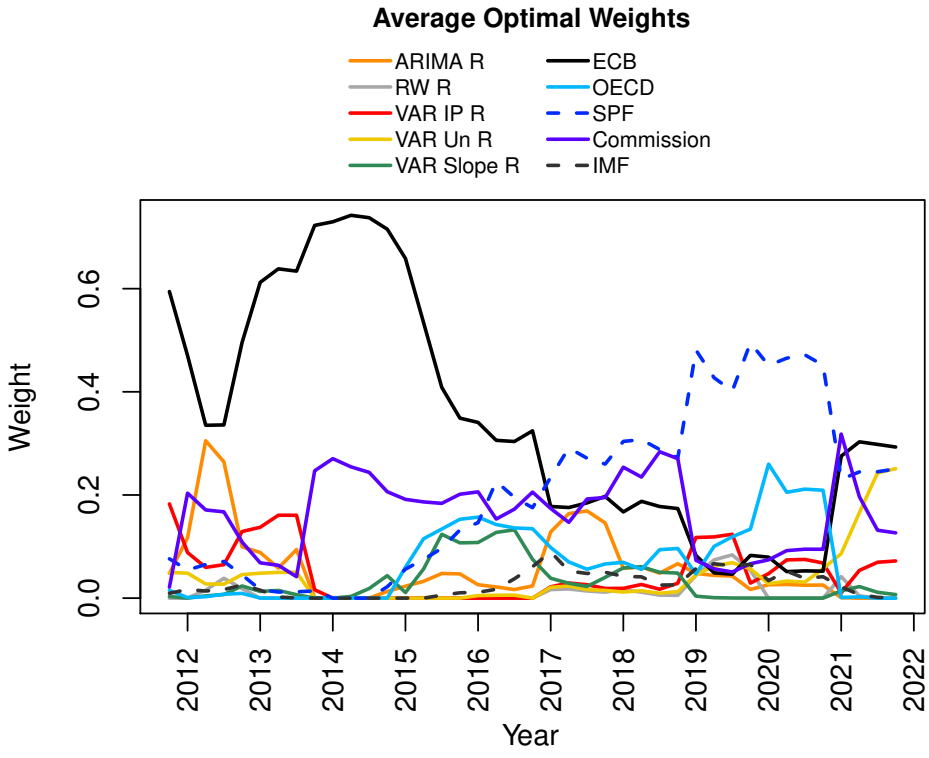


Figure 5.2: Dynamics of average optimal combination weights across specifications.

despite the weights display a more evident instability due to the small sample: in these cases, only ten observations are available to optimally combine ten forecasting methods. Results are available upon request.

Figure 5.2 points out that weights greatly vary over time. Despite the ECB generally receives the greatest weight, its weight is not consistently greater than all remaining forecasts in each sub-period. The weights associated with the other forecasts are also fluctuating over time. These findings underline the importance of the considered reference period and how different conclusions can be drawn when evaluating unconditional forecasts encompassing or conditioning on some economic states. The varying underlying process driving the HICP in the euro area, the heterogeneity of forecasters' loss functions, and heterogeneous information sets can all be plausible explanations for the reason why optimal weights change over time. Therefore, it is compelling to explore whether we can identify these *states* and analyze how the macroeconomic or financial conditions affect the informativeness of the ECB forecasts, and therefore the ability of the ECB at tracking expected inflation.

5.3 The Dynamics of the ECB weight

From the forecast encompassing perspective, the weights reflect the contribution of the forecasts to the forecast that minimizes the variance of the aggregate prediction error. Therefore, a time-varying optimal weight indicates that also the informational value added by the forecast of interest varies over time.

Time-varying weights for model-based forecasts could reflect the uncertainty surrounding the forecasting method as, for example, the uncertainty around the choice of (p, d, q) in ARIMA models, the size of the estimation window or the presence of structural breaks in the relation between the time-series in the bi-variate VARs. These issues are directly connected to the notion of *heterogeneity* of macroeconomic time-series (Clements and Hendry, 1998), i.e., the nonconstant underlying process generating of the series to forecast. Heterogeneity also matters for institutional forecasters. Being supported by econometric models and expert judgment, the predictive performance of institutional forecasters can be affected by the forecasting techniques used and the turnover of forecasters within the same institution.

The loss function of monetary authority has also important consequences on inflation forecasting. For example, Capistrán (2008) argues that the FED systematic under-prediction and over-predictions of inflation can be explained by the variations of the cost of having inflation above an implicit time-varying target, implying the existence of asymmetric loss functions. Furthermore, while all institutional forecasts are certainly useful for consumers and investors, the ECB macroeconomic projections (of which their inflation forecasts are part) are their primary support for the assessment of economic conjecture and risks to price stability which is the primary objective of the ECB mandate. Therefore, we expect the ECB forecasts to have a direct effect on the monetary policy implemented in the euro area: when the expected inflation is low, we expect the monetary authority to try to keep it low to pursue its goal in the medium term. However, policymakers face a dilemma when expected inflation is high: they would like to disinflation, but fear the costs related to a tightening policy (Ball, 1992). These costs do depend on the state of the economy. In sluggish economic growth, high indebtedness and a vulnerable financial system can strongly amplify the negative effects of the recession that would result from such contractionary policies. In that case, the forecasting community does not know how

promptly the monetary authority will react and thus the uncertainty surrounding the inflation dynamics will rise.

Therefore, examining whether the economic environment in which the ECB operates affects the informativeness of its forecasts can offer valuable insights to forecast users looking to enhance their performance, as well as to policy analysts seeking to evaluate the ECB's effectiveness in monitoring inflation. We consider determinants reflecting both the uncertainty surrounding the HICP dynamics and macro-financial factors that may affect the ECB loss functions.

5.3.1 The determinants of the combination weights

Identifying the exact determinants of the weights is not an easy task. We do not directly observe whether (and when) the HICP time series has experienced structural breaks and how the current (and expected) economic conjecture and risks to price stability have shaped forecasters' loss functions. Nevertheless, we can identify a group of variables that might *indirectly* affect the uncertainty around the expected path of the HICP.

5.3.1.1 Disagreement and inflation surprise

Following Mankiw, Reis, and Wolfers (2004) and Manzan (2011), we consider the dispersion of (point) forecasts from the survey of professional forecasters to proxy forecasting uncertainty. Patton and Timmermann (2010) empirically show how this measure mostly reflects the heterogeneity in the priors or underlying forecasting models compared to heterogeneity in information signals. Dovern, Fritsche, and Slacalek (2012) confirm that disagreement about inflation forecasts increases with uncertainty about the actual series, but also find that disagreement about prices rises with their level and, in line with the thesis of Rogoff (1985) and Alesina and Gatti (1995), disagreement declines under independent central banks that promote price stability via the introduction of clear mandates in terms of price stability or the adoption of more predictable monetary policy with increased and improved communication. To study whether these aspects influence the dynamics of the ex-post optimal weights, we include a measure of the difference between the realized y-o-y monthly HICP growth rate and the 2% target (*Difference from 2%*) and the implied dispersion of (point) forecasts from the survey of professional forecasters *Disagreement SPF*. *Disagreement SPF*, is measured as the standard deviation of point forecasts from the ECB Survey of Professional Forecasters.[19] We also control for a broader measure of policy-related economic uncertainty in the euro area, employing the news-based Economic Policy Uncertainty index of Baker, Bloom, and Davis (2016) in our analysis.

Supply shocks, i.e., the sudden change of the supply of a product or commodity, resulting in an unforeseen change in price, e.g., the surge of gas and oil prices following the Ukrainian war, directly affect inflation forecasting performance and, indirectly, the optimal weights. Some forecasters may have an advantage in providing accurate and timely forecasts due to factors such as superior information or lower costs of making revisions. This could enable them to better anticipate or quickly respond to recent economic events

[19]Results are comparable when we considered alternative measures of dispersion for our measure. Specifically, we also considered interquartile or the 95^{th} - 5^{th} range for the *Disagreement SPF*.

(Ehrbeck and Waldmann, 1996). We proxy the dynamics of inflation surprise with the first principal component of realized forecast errors (*Inflation surprise*).

5.3.1.2 Financial conditions and monetary policy

The primary objective of the ECB is to maintain price stability by aiming for 2% inflation over the medium term. However, other objectives can be pursued "without prejudice to the objective of price stability".[20] These objectives include among others the stability of the financial system, the risk of financial fragmentation and the risk of euro break up. In this respect, following the outburst of the European sovereign debt crisis, the ECB has taken a range of actions beyond monitoring price stability to address bank funding problems, eliminate excessive risk in sovereign markets, and safeguard monetary transmission (Gross and Zahner, 2021). These measures have been successful at monitoring these risks when the euro area was experiencing low (or even negative) inflation (Candelon, Luisi, and Roccazzella, 2022; De Grauwe and Ji, 2022). However, the Russian invasion of Ukraine, amid an already slowing recovery from the pandemic, raises new financial stability risks, poses questions about the longer-term impact on economies and the surging commodity prices pose challenging trade-offs for central banks (IMF, 2022a). In the current situation, it is still not clear how the ECB would react when facing such a dilemma.

We analyze whether the optimal weights were affected by measures of financial stability in the euro area. Specifically, we consider the *global interdependence* and *intra-European fragility* factor of Candelon, Luisi, and Roccazzella (2022). The first factor captures shifts in the level of the entire cross-section of long-term spreads, and it is related to factors that capture devaluation risk (De Santis, 2019), flight-to-liquidity mechanisms (Monfort and Renne, 2013) and shifts to global risk aversion (Favero, 2013), further supporting the use of the first PC as a global trend. The second factor measures instead how deep the difference between fragile and financially sounder economies is.[21]

For an exhaustive analysis of the financial conditions in the euro area, we also include proxies mirroring the monetary policy of the ECB. We consider the *Shadow* rate of Wu and Xia (2020) and the *APP Factor* that summarizes Eurosystem holdings corresponding to the ECB Asset Purchase Programme (APP). [22]

5.3.2 Seasonality of institutional forecasts

When evaluating institutional forecasts, the literature often overlooks that forecasts always refer to the average of the calendar year.[23] In this framework, seasonal patterns in the dynamics of the estimated weights might emerge, and ignoring them could result in omitted variable bias when studying the relation with their determinants. Therefore, in

[20] See the ECB website for more details `https://www.ecb.europa.eu/mopo/intro/html/index.en.html`

[21] The factors are extracted using principal component analysis using the samples described in Candelon, Luisi, and Roccazzella (2022).

[22] We compute the factor by extracting the first principal component of the Eurosystem's holdings purchased under one or more of the asset purchase programs operated by the ECB. Specifically, APP includes corporate sector purchase program (CSPP), a public sector purchase program (PSPP), an asset-backed securities purchase program (ABSPP), and a third covered bond purchase program (CBPP3). For data and more information, we refer to `https://www.ecb.europa.eu/mopo/implement/app/html/index.en.html`

[23] In the first quarter of the year, the current-year forecast is four quarters ahead, in the remaining quarters of the year, the horizon becomes three, two, and one quarter(s) ahead, respectively.

the spirit of Lovell (1963) and Saikkonen and Lütkepohl (2000), we complete our set of determinants with three seasonal dummy variables corresponding to the second, third, and fourth quarters of the year. This seasonal adjustment implicitly assumes that the intercept of the regression function shifts each quarter and it is convenient to study how the weights are affected when approaching inflation releasing date.

5.3.3 Controlling for unemployment and industrial production

When studying the dynamics of the weight of the ECB, we should control for quarterly unemployment and quarterly year-on-year industrial production growth rate. While both series are undoubtedly important to represent economic conditions in the euro area, they are also commonly employed in the evaluation of inflation dynamics and monetary policy.

On the one hand, despite monetary policy in the euro area being more complicated than what the conventional Taylor rule might suggest, Gorter, Jacobs, and de Haan (2008) find that the ECB considers *expected*, i.e., forecasted, inflation and output growth when setting interest rates.[24] On the other hand, it is not uncommon to find reformulation of the generalized Phillips curve in terms of industrial production or unemployment rate (Stock and Watson, 1999).

Table 5.4 summarizes the variables, the corresponding abbreviations, and the sources. Before introducing the modeling framework, it is compelling to inspect the dynamics of the determinants themselves to analyze whether potential issues due to multicollinearity may arise.

Table 5.4: Names and abbreviations of variables employed for the analysis of the determinants of the ex-post optimal weights.

Variable	Measure of	Abbreviation	Source
Difference with 2%	Uncertainty	Dif. 2%	Eurostat
Disagreement SPF	Uncertainty	Dis.	ECB
Economic Policy Uncertainty Index	Uncertainty	EPU	Baker, Bloom, and Davis (2016)
Shadow Rate	Monetary Policy	S	Wu and Xia (2020)
Asset Purchasing Program factor	Monetary Policy	APP	ECB
Cross-sectional forecast error factor	Inflation Surprise	Inf. Sur.	Eurostat & ECB
Global Interdependence Factor	Financial Conditions	Glob.	Candelon, Luisi, and Roccazzella (2022)
Fragility Factor	Financial Conditions	Frag.	Candelon, Luisi, and Roccazzella (2022)
Industrial Production	Macroeconomic Conditions	IP	Eurostat
Unemployment rate	Macroeconomic Conditions	Un	Eurostat
Seasonal dummy Q2	Seasonality & Publ. gap	d_{Q2}	·
Seasonal dummy Q3	Seasonality & Publ. gap	d_{Q3}	·
Seasonal dummy Q4	Seasonality & Publ. gap	d_{Q4}	·

Table 5.5 reports the correlation between the determinants in the sample 2021 Q1 - 2021 Q4. We remark that the dynamics of the shadow rate and unemployment rate are strongly positively correlated (over .9). Similarly, the Economic Policy Uncertainty Index is negatively correlated with the shadow rate and unemployment rate ($-.63$ and $-.54$, respectively). As a result, we expect that including the unemployment rate between the determinants, may be detrimental for the precision of the estimated coefficients. We next introduce the regression model used to study the dynamics of ECB's weights.

[24] Contrasting with more conventional specifications of the Taylor rule where realized inflation and output are used.

Table 5.5: Correlation between the determinants in the period Q1 2012 - Q4 2021.

	APP	Un	IP	Dis.	S	Glob.	Frag.	Dif. 2%	Inf. Sur.
EPU	.25	-.54***	-.17	.15	-.63***	.05	.01	.01	-.14
APP		-.11	.18	.01	-.18	.51***	.08	-.34**	-.19
Un			.09	.18	.97***	-.37**	.22	-.10	.14
IP				-.32***	.09	.45***	.12	.15	-.18
Dis.					.03	-.10	.23	-.36***	-.01
S						-.43***	.14	.04	.09
Glob.							-.17	-.47***	.03
Frag.								.09	-.03
Dif. 2%									-.51***

Notes. The marks $***$, $**$ display 1%, and 5% significant level, respectively.

5.3.4 The regression framework

We study the relationship between the weight of the ECB $\boldsymbol{w^{ECB}}$ and their determinants $\boldsymbol{X}$. Assuming $w_t^{ECB} \sim \mathcal{N}(\boldsymbol{x}_t\beta, \sigma)$, we can estimate the following linear regression model:

$$\boldsymbol{w^{ECB}} = \boldsymbol{X\beta} + \boldsymbol{u} \; . \tag{5.14}$$

Combination weights are, nevertheless, defined in the closed interval $[0, 1]$, therefore predictions arising from a linear regression framework may lead to negative weights or weights greater than unity. Despite it is possible to adapt the standard linear regression framework to accommodate bounded stochastic processes by transforming the original weights to have values in the real line[25], this approach has two main shortcomings. First, the regression parameters are interpretable only in terms of the mean of the transformed weights. Second, data generated by bounded stochastic processes are generally skewed and prone to heteroskedasticity.

As a robustness check, we also consider a beta regression model, which is instead tailored to model variables in the interval $(0, 1)$.[26] Specifically, we employ the beta regression framework proposed by Ferrari and Cribari-Neto (2004) that assumes a) independent realizations of $w_i^{ECB} \sim \mathscr{B}(\mu_i, \phi_i)$, b) a linear regression model for the transform of the mean parameter μ and c) a constant precision parameter ϕ. In this framework, we consider a *logit* link function for the mean of w^{ECB}, leading to

$$\mu_i = \frac{e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}} \; . \tag{5.15}$$

We refer to Appendix A5.3 for more details on the beta regression framework.

To tackle the problems arising from potential heteroscedastic and autocorrelated residuals, we report four-lag Newey-West (NW) standard errors (Newey and West, 1987) and a wild bootstrap procedure as in Mammen (1993) both in the linear and beta regression framework.[27]

[25] For example via standard *logit* or Box's transformations.

[26] Following Smithson and Verkuilen (2006), we consider the transformation $\frac{w^{ECB}(n-1)+0.5}{n}$ where n is the sample size to adapt the beta regression framework to dependent variables assuming also the extremes 0 and 1.

[27] We refer to Appendix A5.5 for a detailed description of the bootstrap algorithm.

5.3.5 Empirical results

In Table 5.6, we present the results on ECB's determinants obtained using linear regression and beta regression estimated via OLS and maximum likelihood, respectively. We report the results from the optimal weights obtained in *Specification 2* as the POET estimator of the covariance matrix (Fan, Liao, and Mincheva, 2013) offers better small sample properties compared to the sample covariance matrix.[28]

In panel *a)* we report the main results, while in panel *b)* we also control for unemployment and industrial production. Table 5.6 displays the estimated coefficients, their standard errors, and the p-values implied by testing whether the coefficient of interest is statistically different from zero. Columns NW and B report whether the coefficient of interest is statistically significant at 1% (***), 5% (**), 10% (*) levels according to the Newey-West standard errors and the Mammen wild bootstrapped quantiles, respectively.

We find that the measures of uncertainty (*Dis.* and *EPU*), how actual inflation differs from the 2% target (*Dif. 2%*) and *Inflation Surprise* and the *Global Inter.* are all important drivers of the weight of the ECB. They are statistically significant at 5% level also when controlling for unemployment and industrial. Ceteris paribus, the greater these factors, the lower the expected weight.

The determinants referring to the ECB monetary policy, i.e., the APP factors (*APP*) and the shadow rate (*S*) influence the informativeness of the ECB interestingly: ceteris paribus, the greater the shadow rate, the greater the weight and the greater the APP factor, the lower the weight. This can be expected: higher interest rates signal the intention of central banks to slow down inflation dynamics; while, on the contrary, increasing the amount of holdings purchased under the Asset Purchase Programme and the corresponding liquidity injections decrease ECB's weight. The significance of the ECB monetary policy to the informativeness of its forecasts is challenged when controlling for unemployment and industrial production. However, this issue must be expected: despite the estimated coefficients do not vary substantially, the precision of the estimators is strongly affected by the multicollinearity with unemployment and industrial production as noticed in subsection Table 5.5. Indeed, both *IP* and *UnEmp* are not statistically different from zero at any significance level. We also notice that including *IP* and *UnEmp* does not increase the explanatory power of the model (similar adjusted and pseudo R^2).

5.3.6 Discussion: forecast informativeness and monetary policy

In our framework, weights reflect how informative forecasts are: an upward (downward) dynamics of the weight signals that the considered forecaster has become more (less) informative when producing inflation forecasts. Being more informative implies that the public will regard an institution featuring a greater weight as more relevant when forecasting the variable of interest. In subsection 5.3.5, the greater the difference between current inflation and 2% target, the greater the inflation surprise, and the less informative the ECB forecasts are.

Clearly, as a strategy for conducting monetary policy, having informative forecasts is crucial. First, since realized inflation typically responds to changes in monetary

[28] We report the results for the sample and shrinkage estimator of the covariance matrix in Appendix A5.6.

Table 5.6: Determinants of the weights constructed using the POET estimator of the covariance matrix of prediction errors.

	a) Determinants						b) Determinants + Un and IP					
	LM	NW	B	Beta	NW	B	LM	NW	B	Beta	NW	B
b_0	0.879	***	***	2.090	***	***	0.241			-2.296		
	(0.08)			(0.44)			(0.54)			(3.01)		
Dif. 2%	-0.147	***	***	-0.622	***	***	-0.130	***	***	-0.598	***	***
	(0.03)			(0.14)			(0.03)			(0.13)		
Dis.	-0.187	***	***	-1.225	***	***	-0.237	***	***	-1.466	***	***
	(0.04)			(0.25)			(0.07)			(0.26)		
EPU	-0.001	**	***	-0.004	**	**	-0.001	**	***	-0.006	***	*
	(0.0004)			(0.0019)			(0.0005)			(0.0022)		
S	0.063	***	***	0.351	***	***	0.025			0.072		
	(0.01)			(0.04)			(0.03)			(0.18)		
APP	-0.025	*	***	-0.101	*	**	-0.027	*	**	-0.100		*
	(0.01)			(0.06)			(0.01)			(0.06)		
Inf. Sur.	-0.034	***	***	-0.160	***	***	-0.038	***	***	-0.180	***	***
	(0.01)			(0.06)			(0.01)			(0.06)		
Glob.	0.016	**	***	0.106	***	***	0.019			0.072		
	(0.01)			(0.03)			(0.01)			(0.05)		
Fr.	0.019			0.160	**	***	0.015		*	0.113	*	**
	(0.02)			(0.07)			(0.01)			(0.06)		
Un.	.	.	.	.	.	.	0.060			0.392		
							(0.05)			(0.25)		
IP	.	.	.	.	.	.	-0.004			0.018		
							(0.0046)			(0.023)		
d_{Q2}	-0.045			-0.232			-0.047			-0.229		
	(0.03)			(0.16)			(0.03)			(0.16)		
d_{Q3}	-0.060			-0.238			-0.069			-0.226		
	(0.04)			(0.19)			(0.04)			(0.19)		
d_{Q4}	-0.074	**		-0.401	**	*	-0.075			-0.334		
	(0.04)			(0.19)			(0.04)			(0.22)		
Adj. R2	0.84			0.905			0.844			0.903		
N. obs.	38			38			38			38		

Notes. The first columns displays the determinants, b_0 is the intercept. LM and Beta report the estimated coefficients using the linear regression model and the beta regression. While we display Newey-West standard errors in the parenthesis, NW and B report whether the coefficient of interest is statistically significant at 1% (***), 5% (**), 10% (*) level according to the Newey-West standard errors and the Mammen wild bootstrap quantiles, respectively.

policy only with a substantial lag (from one to two years or more), having reliable information about the expected path of price offers a substantial timing advantage when designing monetary policy. Second, Svensson (1997) showed that inflation-targeting also implies expected inflation-targeting: the central bank's inflation forecast becomes an explicit intermediate target. This should simplify both the implementation and monitoring of monetary policy, helping the inflation-targeting central bank and the public to tell whether the monetary policy is indeed fulfilling its mandate to maintain price stability (Bernanke and Mishkin, 1997). However, this could equally have

serious adverse consequences for the central bank's accountability and credibility[29] if the central bank's forecasts were believed to be inconsistent with the expected dynamics of inflation. In this perspective, low weights relative to their historical trend may be interpreted as warning flags, signaling that the ECB's credibility is at stake. As observed in subsection 5.2.2, we find two main dips of the ECB weight: the first in 2012, the second between the beginning of 2019 and the end of 2020 (see Figure 5.2).

In 2012, the ECB announced and implemented a set of conventional and unconventional tools to address the difficulties in the transmission of monetary policy measures, the onset of a credit crunch, and the risk of deflation following the outburst of the European sovereign debt crisis. ECB's effort on conventional and unconventional monetary policies was also motivated by reverting the deflationary trend in the euro area in the period 2014-2018, therefore pushing inflation back to its target.

The period 2019-2021 saw the outbreak of Covid, the most severe shock to the European economy since WWII with significant disruptions to the labor market, and supply chains and posing serious challenges to economic growth and financial stability (IMF, 2022b). Also, the handover from Mario Draghi to Christine Lagarde (November 2019) as the head of the European Central Bank could have brought into question the ECB's new policy direction.[30] Nevertheless, the recent economic and geopolitical developments strongly contrast with the deflationary trend observed in the euro area after the outburst of the European sovereign debt crisis. Despite we can observe the mild rise of the ECB's weight in 2021, inflation dynamics had not been yet impacted by the global supply chain disruption following the surge of gas and oil prices following the recent war in Ukraine. Indeed, despite this event has a direct inflationary effect, it is not in the data since we do not observe the realized yearly inflation rate for 2022. Eurostat's release of the realized year-on-year euro area annual inflation rate in April 2022 shows how severe the inflation shock is: annual inflation was 7.4% in March 2022, up from 5.9% in February. The new release of the Survey of Professional Forecasters points in the same direction: HICP inflation expectations are revised upward to 6% for 2022 and 2.4% for 2023 but unchanged at 1.9% for 2024. Nevertheless, ECB's policy rates remain currently unchanged and, following Governing Council's forward guidance, any adjustments to the key ECB interest rates will be gradual and will take place sometime after the end of the asset purchasing program expected in Q3 2022.

ECB's decision not to promptly adjust its policy to the surge in forecast inflation in 2022 may seem surprising and strongly contrasts with Bank of England and U.S. Federal Reserve policies that instead vowed tough action on inflation, announcing increasing rates until inflation comes under control. However, as underlined in IMF (2022b) or in the

[29] The term "credibility" refers to the degree of confidence that the public has in the central bank's determination and ability to meet its announced objectives.

[30] For example during a press conference on 12 March 2020, Mrs. Lagarde suggested it was the responsibility of governments to protect highly indebted eurozone countries rather than the central bank, triggering public questions on whether the bank would renew the "whatever it takes" policy to protect the eurozone from a recession triggered by the coronavirus outbreak (the verbatim of the speech by Christine Lagarde is available at `https://www.ecb.europa.eu/press/govcdec/mopo/html/index.en.html`). This resulted in a sudden increase in European sovereign bond spreads, leading ECB's president to walk back her previous comments and restate ECB's full commitment to avoid any fragmentation in bond markets.

recent talk of the director general of the Bank of International Settlements[31] ECB's policy is currently on the edge and more restrictive measures are required to take action against the steady increase in price levels. Miranda-Agrippino and Nenova (2022) have shown that ECB's restrictive monetary policy should have a similar effect that the one implemented at the moment in the U.S.; however, in the euro area, it also may create tensions for the banking sector of peripheral countries (Soenen and Vander Vennet, 2022), generating heterogeneity within euro area financial markets, and in particular in the sovereign bond market. At the end of April 2022, sovereign bond spreads relative to the German Bund are rising again in the euro area: over 175 b.p. for Italy, 220 b.p. for Greece, and around 100 b.p. for Spain and Portugal. Fragmentation risk in this market has already reappeared. In this context, the ECB will need to carefully manage the reduction of asset purchases and the implementation of tightening measures to avoid spillover effects that could trigger financial instability within the euro area. At the same time, however, the normalization of monetary policy cannot be ruled out, should price pressure continue to build up.

Notwithstanding the importance of pursuing financial stability and mitigating fragmentation risk, this also brings into question whether aiming for 2% inflation over the medium term remains the *primary* objective, potentially undermining the credibility of the European Central Bank at maintaining price stability.

5.4 Conclusion

This chapter proposes a novel method to evaluate optimal combinations of the different forecasts. It designs a forecast encompassing test, which explicitly considers that weights lay in the unit simplex. This test can be readily extended in situations where the number of forecasts to evaluate greatly exceeds the realizations of the variable of interest. Hence, we evaluate institutions and model-based forecasts of inflation in the euro area between 2012 - 2021.

Although ECB provides the most informative forecast on average, it does not encompass its competitors: SPF and the European Commission also significantly contribute to the minimization of the forecast error variance. We also find that weights greatly vary over time when we re-estimate the forecast combination considering a rolling window.

The dynamics of the weights associated with the ECB inflation forecast are particularly interesting. Even though they appear on average greater than their competitors, they are not consistently remaining at this position in each sub-period. In particular, the weight of the ECB suddenly drops in Q3 2012 and achieved its minimum in Q1 2019. We, therefore, investigate the fundamental factors behind these dynamics. Macro-financial conditions and monetary policy actions explain this variability. The greater the uncertainty surrounding inflation and the difference between the current and the 2% target, the less informative the ECB forecasts are. The more contractionary the monetary policy, the more informative

[31] On April the 5th, Agustin Carsten has indicated that "Central banks need to adjust to this new environment, not least by raising policy rates to more appropriate levels" `https://www.bis.org/speeches/sp220405.htm`.

they are.

The decline of the weight of the ECB and its relation with its determinants are warning flags: the loss of forecast informativeness damages the ECB's leading role in anchoring inflation expectations and might signal that preserving financial stability can be a challenge to the inflation targeting goal in high inflation environments.

APPENDIX

A5.1 The test for forecast encompassing

We aim to design a test for multiple forecast encompassing that explicitly considers the constraints on the estimated weights and that can deal with high dimensional frameworks.

High dimensionality and constrained parameter space create, however, two problems. First, in the presence of high dimensional nuisance parameters, the maximum likelihood estimator is no longer consistent and, despite it being possible to identify some maximum penalized likelihood estimators that are consistent under some conditions, they may not have a tractable limiting distribution even in the fixed dimensional case (Fu and Knight, 2000). This invalidates the standard inferential theory for Wald, Lagrange Multiplier (LM), and Likelihood Ratio (LR) (see for an example Ning and Liu, 2017). Second, the nonnegativity and sum to one are inequality constraints that generate a closed and convex *cone* C.

We follow the decorrelation procedure of Ning and Liu (2017) to handle the impact of high dimensional nuisance parameters. This approach offers two main advantages: a) not relying on post-selection Lee, Sun, Sun, and Taylor (2016), i.e., considering the conditional inference given the event that some covariates have been selected; and b) the possibility of easily extending the testing procedure under inequality constraints. In fact, following Yu, Gupta, and Kolar (2019), we adapt the decorrelated score function of Ning and Liu (2017) to extend the test for forecast encompassing of low dimensional parameters to a high dimensional framework.

A5.1.1 The testing procedure

We now introduce the test for forecasts encompassing. To ease the presentation, we split the testing procedure into two steps. In **Algorithm 1**, we estimate the ex post optimal combination of forecasts and estimate the decorrelated score function, parameter(s) of interest, and likelihood function. In **Algorithm 2**, we form the corresponding test statistics, compute tail probabilities under the null hypothesis and finally perform inference on the combing weight(s).

The key step in **Algorithm 1** is to estimate ex post optimal weights and the decorrelation operator W. Since both the weights and the nuisance score functions can be high dimensional (have dimension d and d-1, respectively), we need to impose some assumptions on $\boldsymbol{\beta}$ and W to bound the estimation error when $d > n$. On the one hand, $\boldsymbol{\beta}$ searches for the linear combination of weights that minimizes the variance of the aggregate forecasting error under the nonnegativity and sum to one constraints. When $d > n$, we impose $\boldsymbol{\beta}$ to be sparse. Similarly, when $d > n$, W searches for the best sparse linear combination of the nuisance score functions to approximate the score function.

The key step in **Algorithm 2** is to compute the cumulative tail probability of the considered test statistics for a given significance level a. To do this, we have to compute the chi-bar-squared weights. when the constrained space is the nonnegative orthant (or more generally when the constraints defining C are linear and independent), Silvapulle and Sen (2004) propose a simple quadratic programming problem to estimate ω (see proposition 3.6.1).When θ is univariate ($r = 1$), the weights are known in analytical form

Algorithm 1: Decorrelation procedure for high dimensional regression frameworks

1 Estimate ex post optimal weights
$\widehat{\boldsymbol{\beta}}_{\setminus d} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{+(d-1)}} \left\| V_d - \boldsymbol{V}_{\setminus d}\boldsymbol{\beta}_{\setminus d}\right\|_2^2 , \quad \widehat{\beta}_d = 1 - \sum_{i=1}^{d-1} \widehat{\beta}_{\setminus d,i} ,$

2 Define the partition $\beta = (\theta, \gamma)$

3 Estimate the decorrelation operator W_j
$\widehat{W}_j = \arg\min_{W_j\in\mathbb{R}^{d-1}} \left\| V_j - \boldsymbol{V}_{\setminus j} W_j\right\|_2^2 + \lambda \left\| W_j \right\|_1 ,$

4 When θ is a r-low-dimensional multi-dimensional parameter of interest corresponding to the forecasts in the set I, W can be solved column by column i.e., $\widehat{\boldsymbol{W}} = \left(\widehat{W}_j, \text{ where } j \in I\right)$

5 Estimate the decorrelated score function:

$$\widetilde{S}(\theta) = \nabla_\theta \ell(\theta, \widehat{\boldsymbol{\gamma}}) - \boldsymbol{W}^T \nabla_\gamma \ell(\theta, \widehat{\boldsymbol{\gamma}}) ,$$

6 Estimate the decorrelated parameter of interest:

$$\widetilde{\theta} = \widehat{\theta} - \widetilde{\mathscr{F}}_{\theta|\gamma}^{-1}\widetilde{S}, \quad \text{where} \quad \widetilde{\mathscr{F}}_{\theta|\gamma} = \nabla^2_{\theta\theta}\ell(\widehat{\boldsymbol{\beta}}) - \boldsymbol{W}^T\nabla^2_{\theta\gamma}\ell(\widehat{\boldsymbol{\beta}}) ,$$

7 Estimate the decorrelated likelihood function:

$$\widetilde{\ell}(\boldsymbol{\theta}) = \ell(\theta, \widehat{\boldsymbol{\gamma}} - \boldsymbol{W}^T(\theta - \widehat{\theta})) .$$

Algorithm 2: One-sided test for forecasts encompassing

1 Form the decorrelated score test statistics:
$T_S = \left[\widetilde{S}(\widehat{\theta}_{H_0}) - \widetilde{S}(\widehat{\theta}_{H_1})\right]^T \widetilde{\mathscr{F}}_{\theta|\gamma}^{-1} \left[\widetilde{S}(\widehat{\theta}_{H_0}) - \widetilde{S}(\widehat{\theta}_{H_1})\right] .$

2 Form the decorrelated likelihood ratio (LR) test statistics:
$T_{LR} = 2n\left(\widetilde{\ell}(\widehat{\boldsymbol{\beta}}_{H_0}) - \widetilde{\ell}(\widehat{\boldsymbol{\beta}}_{H_1})\right) ,$

3 Form the decorrelated Wald test statistics:
$T_W = \left[\widetilde{\theta} - \widehat{\theta}_{H_0}\right]^T \widetilde{\mathscr{F}}_{\theta|\gamma} \left[\widetilde{\theta} - \widehat{\theta}_{H_0}\right] - \left[\widetilde{\theta} - \widehat{\theta}_{H_1}\right]^T \widetilde{\mathscr{F}}_{\theta|\gamma} \left[\widetilde{\theta} - \widehat{\theta}_{H_1}\right] .$

4 Compute the set of mixing weights $\omega_i(r, \widetilde{\mathscr{F}}_{\theta|\gamma}, \mathbb{R}^{+r})$ for the $\overline{\chi}^2$ statistics.

5 Given a significance level a and its corresponding critical value c_a, compute cumulative tail probability :

$$\text{Prob}\left[\overline{\chi}^2(\widetilde{\mathscr{F}}_{\theta|\gamma}, \mathbb{R}^{+r}) > c_a\right] = \sum_{i=0}^{r} \omega_i \text{Prob}\left[\chi^2(i) > c_a\right] .$$

6 Reject H_0 if

$$\text{Prob}\left[T_S > c\right] > a , \quad \text{Prob}\left[T_W > c\right] > a, \quad \text{Prob}\left[T_{LR} > c\right] > a .$$

and the chi-bar-squared statistics reduce to

$$\text{Prob}\left[\overline{\chi}^2(\widetilde{\mathscr{F}}_{\theta|\gamma}, \mathbb{R}^{+r}) > c_a\right] = .5 \text{ Prob}\left[\chi^2(0) > c_a\right] + .5 \text{ Prob}\left[\chi^2(1) > c_a\right] . \qquad \text{(A5.1)}$$

To ease presentation and simplify notation, we focus on the likelihood framework and on the univariate parameter of interest θ ($r = 1$) corresponding to the j^{th} forecast. Nevertheless, the decorrelated score function can be defined for a multi-dimensional parameter of interest and also when $\boldsymbol{\beta}$ minimizes a loss function other than the negative log-likelihood. (see section 3 and supplementary material of Ning and Liu, 2017, respectively).

A5.2 Assumptions of the test when $d > n$

In this section, we describe the high-level assumptions required to obtain weak convergence of test under null hypothesis in the high dimensional framework. We closely follow Ning and Liu (2017) and Yu, Gupta, and Kolar (2019) and refer to their papers for proof of the results in the case of linear regression models. These conditions can be classified into three main categories: (a) Consistency conditions for initial parameter estimation (Assumption 1, 2, and 3); (b) Concentration of the gradient and Hessian matrix (Assumption 4); (c) Local smoothness on the loss function (Assumption 5). All these assumptions are fairly standard in the high dimensional statistics literature and are known to hold for the Lasso and nonnegative least square with sum to one constraint on the weights (Fan, Zhang, and Yu, 2012).

Assumption A5.2.1. *Score condition. The expected value of the score function at the true $\boldsymbol{\beta}^*$ is equal to zero, i.e., $\nabla \ell(\boldsymbol{\beta}^*) = 0$.*

Assumption A5.2.2. *Restricted eigenvalue condition for the covariance matrix. Given the positive definite matrix $\mathscr{F}$ and $\nabla \ell^2(\boldsymbol{\beta})$, let $\mathcal{S} = supp(\boldsymbol{\beta}^*) \cup supp(\boldsymbol{W}^*)$. There exists a universal constant $c, k > 0$ such that the quadratic form $\boldsymbol{v}^t \mathscr{F} \boldsymbol{v} \geq k\|\boldsymbol{v}\|_2^2$, and $\boldsymbol{v}^T \nabla \ell^2(\boldsymbol{\beta}) \boldsymbol{v}$, for any $\boldsymbol{v}$ in the cone C, i.e., $\forall \boldsymbol{v} \in C(\mathcal{S}) = \{\boldsymbol{v} : \|\boldsymbol{v}_{\mathcal{S}^c} \leq c\boldsymbol{v}_{\mathcal{S}}\|\}$.*

Assumption A5.2.3. *Consistency conditions for initial parameter estimation. As $n \longrightarrow \infty$, for some sequences $\nu_1(n)$ and $\nu_2(n)$ converging to 0, the following holds*

$$\lim_{n\to\infty} Prob_{\boldsymbol{\beta}^*}\left[\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \lesssim \nu_1(n)\right] = 1 \ , \quad and \quad \lim_{n\to\infty} Prob_{\boldsymbol{\beta}^*}\left[\|\widehat{\boldsymbol{W}} - \boldsymbol{W}^*\|_1 \lesssim \nu_2(n)\right] = 1 \ .$$

In lemma 3.1 of the paper, D3 and D4 of the supplementary material, Ning and Liu (2017) show that this assumption holds when considering the Lasso estimator of $\boldsymbol{\beta}$ and $\boldsymbol{W}$ in a linear regression model. In this case with high probability, we have the following

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 = \mathcal{O}\left(s^* \sqrt{\frac{log(d)}{n}}\right) \ , \quad \|\widehat{\boldsymbol{W}}_j - \boldsymbol{W}_j^*\|_1 = \mathcal{O}\left(s^{'} \sqrt{\frac{log(d)}{n}}\right) \text{ with } j \in I \ ,$$

where $\boldsymbol{\beta}$ and $\boldsymbol{W}$ are defined in **Algorithm 1** and $s^* = \|\boldsymbol{\beta}^*\|_0$ and $s^{'} = \|\boldsymbol{W}^*\|_0$, i.e., the number of non zero elements in $\boldsymbol{\beta}^*$ and $\boldsymbol{W}^*$, respectively.[32]

Assumption A5.2.4. *Concentration of the gradient and Hessian. Let $v^* = (1, -\boldsymbol{W}^{T*})$ and assume that the largest element of the score function is bounded, i.e., $\|\nabla \ell(\boldsymbol{\beta}^*)\|_\infty = \mathcal{O}\left(\sqrt{\frac{log(d)}{n}}\right)$ and*

$$\|v^{T*} \nabla^2 \ell(\boldsymbol{\beta}^* - \mathbb{E}_{\boldsymbol{\beta}^*}\left[v^{T*} \nabla^2 \ell(\boldsymbol{\beta}^*\right] \|_\infty = \mathcal{O}\left(\sqrt{\frac{log(d)}{n}}\right) \ .$$

[32] Remark that $\boldsymbol{W}^* = \mathscr{F}_{\gamma\gamma}^{*-1} \mathscr{F}_{\gamma\theta}^*$.

Assumption A5.2.5. *Local smoothness of the loss function For some constant L, the Hessian matrix $\nabla^2\ell(\boldsymbol{\beta})$ is Lipschitz continuous:*

$$\|\nabla^2\ell(\boldsymbol{\beta}_1) - \nabla^2\ell(\boldsymbol{\beta}_2)\|_\infty \leq L\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_1 \ .$$

Lemma A5.2.1. *Central limit theorem (CLT) for the linear combination of the score function. Let $\boldsymbol{\Sigma}^* = \lim_{n\to\infty} Var_{\boldsymbol{\beta}^*}[\sqrt{n}\nabla\ell(\boldsymbol{\beta}^*)]$. Under **Assumption 1-5**, Ning and Liu (2017) and Yu, Gupta, and Kolar (2019) proved that*[33]

$$\sqrt{n}\boldsymbol{v}^{T*}\nabla\ell(\boldsymbol{\beta}^*)\sqrt{\sigma^*}^{-1} \rightsquigarrow \mathcal{N}(0,1), \quad \textit{where } \sigma^* = \boldsymbol{v}^{T*}\boldsymbol{\Sigma}^*\boldsymbol{v}^* \ .$$

A5.3 Beta regression framework

The density function of a beta distribution is given by

$$f(y,p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\ \Gamma(q)}y^{p-1}\ (1-y)^{q-1}, \quad \text{with } 0<y<1\ , \tag{A5.2}$$

with $p,q>0$ and $\Gamma(\cdot)$ being the gamma function. However, the beta distribution can be conveniently re-parameterized in terms of $0<\mu<1$ and $\phi>0$ that are related to the mean and variance of the dependent variable y,

$$E[y_i] = \mu_i \text{ and } Var[y_i] = \frac{\mu_i(1-\mu_i)}{1+\phi}\ . \tag{A5.3}$$

In this framework, the log-likelihood function becomes

$$\ell(\mu,\phi) = \sum_{i=1}^{n}\ell_i(\mu_i,\phi)\ , \tag{A5.4}$$

where

$$\begin{aligned}\ell_i(\mu_i,\phi) = log\left[\Gamma(\phi)\right] - log\left[\Gamma(\mu_i\phi)\right] - log\left\{\Gamma\left[(1-\mu_i)\phi\right]\right\} + \\ + (\mu_i\phi-1)\,log(y_i) + \left[(1-\mu_i)\,\phi-1\right]log(1)y_i)\ .\end{aligned}$$

Using this notation, the standard beta regression model assumes a) independent realizations of $y_i \sim \mathscr{B}(\mu_i,\phi_i)$, b) a linear regression model for the transform of the mean parameter and c) a constant precision parameter ϕ. Fixing the precision parameter makes the beta regression framework more parsimonious at the cost of lower flexibility than the variable dispersion beta regression models featured in Gambetti, Gauthier, and Vrins (2019). However, this is a price we have to pay considering that we count only 38 observations in our sample. Nevertheless, we can easily notice that this model can still naturally accommodate heteroscedasticity because $Var[y]$ is also a function of μ_i. Formally,

$$g_1(\mu_i) = x_i^T\beta \text{ and } g_2(\phi_i) = \phi\ , \tag{A5.5}$$

with $g_1(\cdot):(0,1)\to\mathbb{R}$ and $g_2(\cdot):(0,\infty)\to\mathbb{R}$, with x_i being the vector of regressors, β conformable vectors of regression coefficients, and $g_1(\cdot)$ and $g_2(\cdot)$ being strictly increasing and twice-differentiable link functions. We consider a *logit* link function for g_1 and a constant g_2. This results in

$$\mu_i = \frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}} \ \text{ and } \ \phi_i = \phi\ . \tag{A5.6}$$

[33] We refer to Appendix C and D of Ning and Liu (2017) for the proofs that Assumption 1 to 5 hold in a general linear regression model, while we refer to the appendix of Fan, Zhang, and Yu (2012) for the proofs related to the nonnegative least squares estimator under sum to one constraint.

A5.4 On the empirical distribution of ϵ

We analyze the distribution of the residual forecast error, denoted by ϵ, which is implied by the ex-post optimal combinations of forecasts. In section 5.1, we have assumed that, in population, ϵ follows a Normal distribution with mean 0 and standard deviation σ_ϵ. To test whether this assumption holds in our empirical application, we present the descriptive statistics of ϵ in Table A5.1 and compare our empirical distribution to the Normal density using the Kolmogorov-Smirnov test. The p-value of .11 suggests that we cannot reject the null hypothesis of no difference between the two distributions at the usual 90%, 95% and 99% confidence levels.

Table A5.1: Descriptive statistics of the prediction errors of the optimal combinations of forecasts.

	Mean	sd	min	max	range	skew	kurtosis	
	-0.04	0.37	-1.05	1.42	2.47	0.84	4.46	
				Percentiles				
2.50%	5%	10%	25%	50%	75%	90%	95%	97.50%
-0.766	-0.510	-0.408	-0.165	-0.035	0.070	0.217	0.475	0.732
				Kolmogorov–Smirnov test p-value:			0.1195	

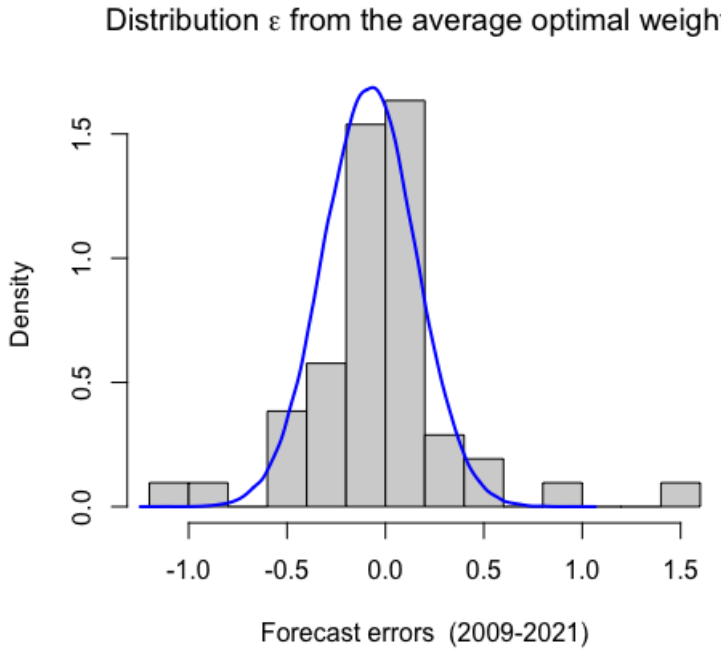


Figure A5.1: Distribution of the forecast errors $\boldsymbol{\epsilon}$ associated to the optimal combination of forecasts estimated in specifications 1, 2 and 3.

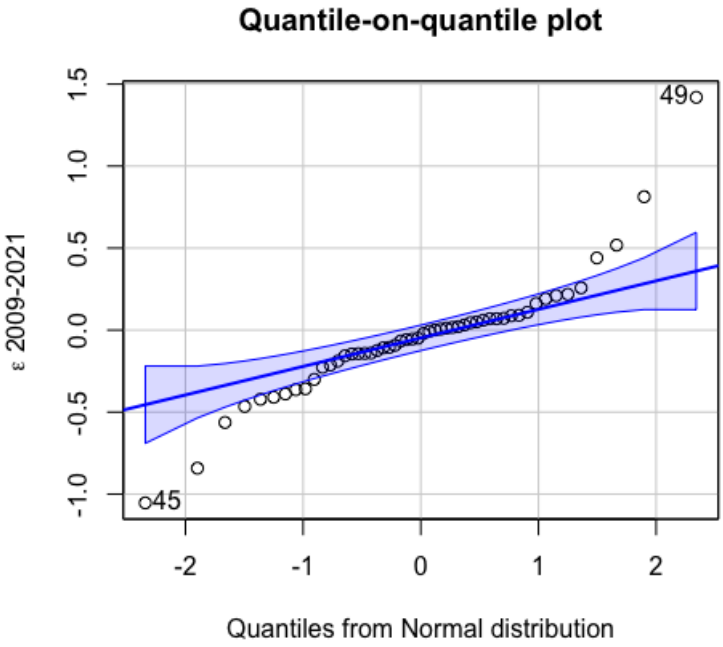


Figure A5.2: Quantile-on-quantile plot of the realized ϵ associated to the average optimal combining weights with the theoretical quantiles from the Normal distribution with 99% confidence bands.

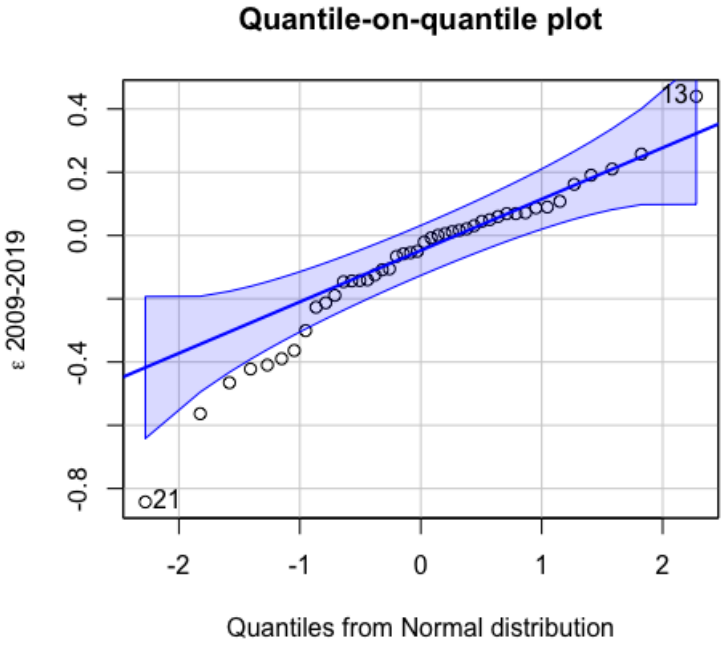


Figure A5.3: Quantile-on-quantile plot of the realized ϵ as in Figure A5.2 but when excluding the Covid period (2020-2021).

We also display the empirical distribution of ϵ in our sample of 52 realizations (2009-2021) in Figure A5.1 and compare it with the Normal density. The distribution is centered around 0 with no apparent skewness, but there is some evidence of leptokurtic tails. To further investigate this aspect, we present a quantile-on-quantile (QQ) plot in Figure A5.2, which compares the empirical quantiles of ϵ with the theoretical quantiles of the Normal density. In a sample of N observations, pointwise confidence bands for the sample $p-$quantile q^p are built from the Normal distribution centered around q^p and with variance $\frac{p(1-p)}{N\ \left[\phi\left(\Phi^{-1}(p)\right)\right]^2}$, with ϕ and Φ being the pdf and the cdf of a standard Normal variable, respectively (Hyndman and Fan, 1996).

The QQ plot indicates that the fat tails are mainly due to observations in the years 2020 and 2021, particularly in Q1 (observations 45 and 49), which coincide with the inflation forecasts made at the beginning of the Covid pandemic and the first phase of the Russian invasion of Ukraine. To address this issue, we present a QQ plot for the sample excluding 2020 and 2021 in Figure A5.3. In this case, the theoretical and empirical quantiles are closer to each other, and the concerns related to leptokurtic tails become, therefore, less relevant.

A5.5 Bootstrap procedure

In our framework, we consider the bootstrapping procedure to take into account also unknown forms of heteroscedasticity that could emerge in the dynamics of the estimated weights, we proposed the wild bootstrapped procedure as in Mammen (1993). Specifically,

1. Estimate the standard linear regression or beta regression parameters via maximum likelihood as described in the subsection 5.3.4 (after retrieving the ex post optimal combination weights as outlined in Section 5.1), and the corresponding residuals u_t^i for each $i = 1, \ldots, n$.
2. Draw with replacement the sequence of errors $\{\tilde{u}_i = k_i \hat{u}_i\}_{i=1}^n$ with

$$k_i = \begin{cases} \frac{1+\sqrt{5}}{2}, & \text{with probability } p = \frac{\sqrt{5}-1}{2\sqrt{5}} \\ \frac{1-\sqrt{5}}{2}, & \text{with probability } 1-p. \end{cases}$$

   as in Mammen (1993) with $\{\hat{u}_i\}_{i=1}^n$ being the estimated errors in the beta regression.
3. Generate the bootstrapped dependent variable using $\{\tilde{y}_i = \widehat{y}_i + k_i\}_{i=1}^n$, <ith $\{\hat{y}_i\}_{i=1}^n$ being the fitted values of the dependent variable in the beta regression. We map negative values or weights greater than one of the bootstrapped samples $\{\tilde{y}_i\}_{i=1}^n$ to zero and one, respectively. Following Smithson and Verkuilen (2006), we consider the transformation $\frac{y(n-1)+0.5}{n}$ where y is the dependent variable and n is the sample size to adapt the beta regression framework to dependent variables assuming also the extremes 0 and 1.
4. Estimate the beta regression on the bootstrapped data sample.
5. Repeat 2 - 4 a large number of times (in this work, we repeat the procedure $10,000$ times), and then extract the quantiles needed.

A5.6 Additional results on the determinants of the weights

Table A5.2: Determinants of the weights constructed using the sample (Spec. 1) and the shrinkage estimator (Spec. 3) of the covariance matrix of prediction errors.

	Specification 1			Specification 3		
	LM	NW	B	LM	NW	B
b_0	0.767	***	***	0.820	***	***
	(0.0621)			(0.0720)		
Dif. 2%	-0.055	*	*	-0.147	***	***
	(0.0289)			(0.0319)		
Dis.	-0.064	**	*	-0.079	**	
	(0.0281)			(0.0306)		
EPU	0.001	**	*	-0.001	***	***
	(0.0004)			(0.0004)		
S	0.117	***	***	0.072	***	***
	(0.0055)			(0.0039)		
APP	0.000			-0.019		
	(0.0103)			(0.0115)		
Inf. Sur.	-0.034	**	***	-0.027	**	**
	(0.0147)			(0.0101)		
Glob.	0.019	***	***	0.005		
	(0.0065)			(0.0062)		
Fr.	-0.014			0.000		
	(0.0204)			(0.0195)		
d_{Q2}	-0.077	**	***	-0.032		
	(0.0368)			(0.0307)		
d_{Q3}	-0.062	*	***	-0.036		
	(0.0310)			(0.0302)		
d_{Q4}	-0.134	***	***	-0.050	*	
	(0.0458)			(0.0253)		

Notes. The first columns displays the determinants, b_0 is the intercept. LM report the estimated coefficients using the linear regression model, while we display Newey-West standard errors in the parenthesis, NW and B report whether the coefficient of interest is statistically significant at 1% (***), 5% (**), 10% (*) level according to the Newey-West standard errors and the Mammen wild boostrap quantiles, respectively.

A5.7 Size and power analysis

In this section, we study the performance of the proposed test both in terms of empirical size and power. Similarly to Busetti and Marcucci (2013), we design a Monte Carlo experiment where we model the j^{th} forecast using a single factor model

$$f_i = x_i + \epsilon_{i,j}, \quad \text{for} \quad i = 1, 2, ...n \tag{A5.7}$$

and allow for $d > n$. The factor f_i is an i.i.d. $\mathcal{N}(0,1)$ and $\epsilon_{i,j}$ is an idiosyncratic and potentially auto-correlated error term, i.e., $\epsilon_{i,j} = \rho\epsilon_{i-1,j} + z_j,\ z_j \sim \mathcal{N}(0,\sigma_z)$. The i^{th} realization of the variable of interest is obtained from

$$y_i = \boldsymbol{\beta}^T \boldsymbol{f}_i + u_i \ . \tag{A5.8}$$

The systematic error term u_i permits us to study the behavior of the test even in the presence of auto-correlated residuals, i.e., $u_i = \rho u_{i-1} + z_i,\ z_i \sim \mathcal{N}(0,\sigma_z)$. $\boldsymbol{\beta}$ is the vector weights that lies in the *d-1* dimensional unit simplex. We partition the vector of weights $\boldsymbol{\beta}$ into two sub-vectors, i.e., $\boldsymbol{\beta} = (\theta, \boldsymbol{\gamma})$, θ is the parameter of interest while $\boldsymbol{\gamma}$ is a sparse vector of nuisance parameters with d non-zero components. Indeed, we require $\boldsymbol{\gamma}$ to be sparse to insure the consistency conditions on $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{W}}$ when the problem is high-dimensional.

We consider several schemes to assign the weights to the non-zero element i of $\boldsymbol{\gamma}$: a) equal weight, i.e., $\boldsymbol{\gamma}_i = \frac{1-\theta}{d}$; b) random, i.e., $\boldsymbol{\gamma}_i = \frac{r}{k}$, where $r \sim \mathcal{U}(0,1)$ and k is a normalization constant such that $\theta + \boldsymbol{\gamma} = 1$. To avoid repetition, we report the results for the scheme a) with $m = 20$, $d = 5$, $n = \{40, 80, 160\}$, $\rho = \{0, .25, .5, .75\}$ and $\sigma_z = .5$ and $\sigma_u = .25$. [34]

Given n realizations of the target and m forecasts, Table A5.3 and Table A5.4 report the results of the simulation study for a nominal size of 5% and 10000 repetitions. We confirm the usual results about the size-adjusted power curve. Power increases when $n \uparrow$ and when $\theta \to 1$. We also expect that when $\rho \uparrow$, the size-adjusted power $\downarrow$. Indeed, we have specified the dynamics of the ϵ_j as an autoregressive process of order 1 - AR(1) - and, keeping σ_z constant, the long-term variance of ϵ_j to tend to infinity as $\rho \to 1$.

[34] The analysis on the empirical size and size-adjusted power in schemes a) and b) are comparable. The parameters $\sigma_z = .5$ and $\sigma_u = .25$ are chosen to obtain an average signal-to-noise ratio ranging from 1.60 (when $\rho = .75$) to 1.8 (when $\rho = 0$) across the simulation study for a sample of 1000 observation. Similarly, we report results for $d = 5$, but other tables with additional robustness checks are available upon request.

Table A5.3: Empirical and size-adjusted rejection probabilities for the analysis of the power and size-adjusted power when acknowledging ($\boldsymbol{\beta}^{\geq 0}$ and $LM \sim \bar{\chi}^2$) or overlooking ($\boldsymbol{\beta}$ and $LM\chi^2$) the nonnegativity constraint on the combining weights.

Rejection probability		Distance from the null: 0 - Size	.01	.025	.05	.10	.20	.25
n : 160, ρ : 0								
$\boldsymbol{\beta}^{\geq 0}$ and $LM \sim \bar{\chi}^2$	Size-adj.	.050	.064	.113	.259	.720	.998	1
	Empirical	.054	.068	.118	.271	.732	.998	1
$\boldsymbol{\beta}$ and $LM \sim \chi^2$	Size-adj. power	.050	.059	.094	.229	.676	.997	1
	Empirical	.085	.095	.145	.304	.751	.998	1
n : 160, ρ : .25								
$\boldsymbol{\beta}^{\geq 0}$ and $LM \sim \bar{\chi}^2$	Size-adj.	.050	.059	.102	.249	.697	.997	1
	Empirical	.054	.064	.108	.258	.708	.997	1
$\boldsymbol{\beta}$ and $LM \sim \chi^2$	Size-adj.	.088	.089	.136	.293	.732	.998	1
	Empirical	.050	.054	.085	.216	.643	.994	1
n : 160, ρ : .50								
$\boldsymbol{\beta}^{\geq 0}$ and $LM \sim \bar{\chi}^2$	Size-adj.	.050	.060	.089	.209	.606	.989	.999
	Empirical	.054	.064	.096	.219	.618	.990	.999
$\boldsymbol{\beta}$ and $LM \sim \chi^2$	Size-adj.	.086	.095	.120	.252	.643	.991	1
	Empirical	.050	.058	.080	.185	.559	.983	.999
n : 160, ρ : .75								
$\boldsymbol{\beta}^{\geq 0}$ and $LM \sim \bar{\chi}^2$	Size-adj.	.050	.054	.080	.149	.411	.914	.983
	Empirical	.054	.060	.086	.157	.425	.919	.984
$\boldsymbol{\beta}$ and $LM \sim \chi^2$	Size-adj.	.082	.088	.112	.186	.456	.926	.985
	Empirical	.050	.053	.069	.128	.368	.889	.975
n : 80, ρ : 0								
$\boldsymbol{\beta}^{\geq 0}$ and $LM \sim \bar{\chi}^2$	Size-adj. power	.050	.058	.086	.170	.435	.929	.987
	Empirical	.057	.068	.098	.185	.459	.936	.989
$\boldsymbol{\beta}$ and $LM \sim \chi^2$	Size-adj.	.130	.138	.166	.254	.528	.942	.990
	Empirical	.050	.051	.068	.127	.351	.875	.971
n : 80, ρ : .25								
$\boldsymbol{\beta}^{\geq 0}$ and $LM \sim \bar{\chi}^2$	Size-adj.	.050	.053	.082	.154	.416	.916	.985
	Empirical	.059	.064	.095	.175	.446	.927	.987
$\boldsymbol{\beta}$ and $LM \sim \chi^2$	Size-adj.	.132	.138	.160	.245	.514	.938	.988
	Empirical	.050	.049	.068	.118	.318	.854	.960

Notes. We estimate the empirical size by generating the optimal forecast under the null hypothesis. We conduct the power analysis by generating the optimal forecast under the alternative for a departure from the null of $\beta = 0$, i.e., by imposing the population weight of the forecast to be tested equal to .01, .025, .05, .10, .20, and .25.

Table A5.4: Empirical and size-adjusted rejection probabilities for the analysis of the power and size-adjusted power when acknowledging ($\boldsymbol{\beta}^{\geq 0}$ and $LM \sim \bar{\chi}^2$) or overlooking ($\boldsymbol{\beta}$ and $LM\chi^2$) the nonnegativity constraint on the combining weights.

Rejection probability		Distance from the null: 0 - Size	.01	.025	.05	.10	.20	.25
$n : 80$, $\rho : .50$								
$\boldsymbol{\beta}^{\geq 0}$ and $LM \sim \bar{\chi}^2$	Size-adj.	.050	.055	.073	.137	.354	.848	.958
	Empirical	.059	.064	.083	.152	.375	.862	.963
$\boldsymbol{\beta}$ and $LM \sim \chi^2$	Size-adj.	.129	.130	.149	.216	.442	.881	.967
	Empirical	.050	.053	.063	.108	.278	.772	.923
$n : 80$, $\rho : .75$								
$\boldsymbol{\beta}^{\geq 0}$ and $LM \sim \bar{\chi}^2$	Size-adj.	.050	.053	.065	.105	.245	.664	.822
	Empirical	.060	.063	.078	.121	.272	.688	.841
$\boldsymbol{\beta}$ and $LM \sim \chi^2$	Size-adj.	.126	.134	.142	.186	.340	.728	.858
	Empirical	.050	.053	.059	.088	.191	.573	.746
$n : 40$, $\rho : 0$								
$\boldsymbol{\beta}^{\geq 0}$ and $LM \sim \bar{\chi}^2$	Size-adj.	.050	.059	.076	.122	.258	.677	.835
	Empirical	.067	.080	.097	.149	.297	.714	.860
$\boldsymbol{\beta}$ and $LM \sim \chi^2$	Size-adj.	.284	.283	.291	.334	.467	.781	.885
	Empirical	.050	.053	.054	.075	.131	.430	.608
$n : 40$, $\rho : .25$								
$\boldsymbol{\beta}^{\geq 0}$ and $LM \sim \bar{\chi}^2$	Size-adj. power	.050	.056	.068	.109	.244	.632	.806
	Empirical	.071	.079	.094	.145	.292	.684	.843
$\boldsymbol{\beta}$ and $LM \sim \chi^2$	Size-adj. power	.284	.283	.291	.330	.449	.753	.869
	Empirical	.050	.050	.052	.068	.128	.398	.559
$n : 40$, $\rho : .50$								
$\boldsymbol{\beta}^{\geq 0}$ and $LM \sim \bar{\chi}^2$	Size-adj.	.050	.060	.074	.108	.226	.585	.744
	Empirical	.066	.076	.090	.126	.257	.618	.772
$\boldsymbol{\beta}$ and $LM \sim \chi^2$	Size-adj.	.274	.279	.282	.310	.420	.708	.821
	Empirical	.050	.051	.052	.068	.121	.356	.508
$n : 40$, $\rho : .75$								
$\boldsymbol{\beta}^{\geq 0}$ and $LM \sim \bar{\chi}^2$	Size-adj.	.050	.059	.059	.083	.163	.394	.539
	Empirical	.069	.079	.080	.109	.199	.445	.586
$\boldsymbol{\beta}$ and $LM \sim \chi^2$	Size-adj.	.258	.266	.261	.278	.362	.562	.669
	Empirical	.050	.051	.052	.061	.098	.238	.344

Notes. We estimate the empirical size by generating the optimal forecast under the null hypothesis. We conduct the power analysis by generating the optimal forecast under the alternative for a departure from the null of $\beta = 0$, i.e., by imposing the population weight of the forecast to be tested equal to .01, .025, .05, .10, .20, and .25.

Table A5.5: Empirical and size-adjusted rejection probabilities for the analysis of the empirical size and size-adjusted power in a large sample and in a high dimensional framework

Rejection probability		Distance from the null: 0 - Size	.01	.025	.05	.10	.20	.25
$n : 1000$, $d : 20$, $\rho : 0$								
$\boldsymbol{\beta}^{\geq 0}$ and $LM \sim \bar{\chi}^2$	Size-adj.	0.050	0.100	0.370	0.898	1.000	1.000	1.000
	Empirical	0.051	0.103	0.374	0.900	1.000	1.000	1.000
$\boldsymbol{\beta}$ and $LM \sim \chi^2$	Size-adj.	0.050	0.097	0.362	0.894	1.000	1.000	1.000
	Empirical	0.055	0.104	0.380	0.902	1.000	1.000	1.000
$n : 40$, $d : 80$, $\rho : 0$								
$\boldsymbol{\beta}^{\geq 0}$ and $LM \sim \bar{\chi}^2$	Size-adj.	0.050	0.047	0.051	0.071	0.127	0.341	0.461
	Empirical	0.054	0.051	0.058	0.076	0.138	0.360	0.479
$\boldsymbol{\beta}$ and $LM \sim \chi^2$	Size-adj.	0.050	0.046	0.051	0.057	0.096	0.242	0.349
	Empirical	0.035	0.033	0.037	0.041	0.071	0.198	0.290

Notes. We study the convergence of the empirical to the theoretical 5% size for a sample size of 1000 observations and the behavior of the proposed test in the high dimensional framework with 40 observations and 80 candidate forecasts. In this latter case, we estimate the unconstrained weights using Lasso. The increase of the sample size was necessary to estimate the penalty intensity via 5-fold cross-validation. We estimate the empirical size by generating the optimal forecast under the null hypothesis, i.e., by enforcing the weight to be tested to be zero in the population. We conduct the power analysis by generating the optimal forecast under the alternative for a departure from the null of $\beta = 0$, i.e., by imposing the population weight of the forecast to be tested equal to .01, .025, .05, .10, .20, and .25.

Chapter 6

Fragmentation in the European Monetary Union: Is it really over?

with Bertrand Candelon and Angelo Luisi

Abstract

Sovereign bond market fragmentation represents one of the major challenges European authorities have had to tackle since the outburst of the euro area debt crisis in 2010. By investigating the inter-country shock transmission through a new methodology that reconciles Factor and Global Vector Autoregressive models, we first show that fragmentation risk well preceded the sovereign debt crisis outburst. Most importantly, by analyzing the recent period, we document a rise in fragmentation risk in the euro area during the COVID pandemic. This rise, connected to the pressure on public debts and deficits due to the pandemic period, questions the European integration process and calls for early measures to avoid a new sovereign debt crisis.

6.1 Introduction

Since the 2010-2012 European Sovereign Debt (ESD) crisis, the evolution of sovereign bonds in the euro area countries has been closely monitored. With the introduction of the euro as a common currency, long-term yield differentials between euro area government bonds started to co-move synchronously at low rates, respecting hence the well-known property of uncovered interest rate parity in fixed exchange rate regimes. The ESD crisis completely changed this setting by leaving place for heterogeneous and unprecedentedly high levels of divergence among yields. Specifically, high- and low-indebted countries exhibited different patterns. The former experienced increasing yields, the latter decreasing. The fragmentation era began.

De Grauwe and Ji (2012) proved how, during the ESD crisis, only Greece experienced a solvency crisis (i.e., driven by underlying fundamentals), while the other high-indebted countries were instead hit by self-fulfilling liquidity crises (i.e., not based on changes in underlying fundamentals). Furthermore, existing literature outlined how during periods of market distress, euro area sovereign bond market participants tend to (i) dis-invest in bonds of countries exhibiting high debt- and deficit-to-GDP ratios (i.e., a temporary shift in preferences, labeled as "contagion", see, for example, Favero and Missale, 2012), and (ii) invest more in safer and more liquid assets (i.e., "flight-to-quality", see, for example, Beber, Brandt, and Kavajecz, 2009).

It, therefore, appears evident that, even during such a period of strong yields' co-movement as the pre-ESD crisis, euro area sovereign bond markets were not fully integrated.

In fact, bond markets can be considered integrated when bonds are exposed to the same risks (Baele, Ferrando, Hördahl, Krylova, and Monnet, 2004). The observed heterogeneous reaction to the asymmetric Greek solvency crisis proved that fragmentation risk (i.e., countries sovereign bond market heterogeneous exposure to foreign counterparts) preceded the ESD crisis outburst (Candelon and Luisi, 2020).

In this chapter, we thus aim at evaluating the presence of a latent fragmentation risk already in the period preceding the ESD crisis outburst. Fragmentation risk is defined in an *out-of-sample* fashion, that is, the risk that a sudden increase in a high-indebted country's yield is associated with (i) an increase in sovereign yields of other high-indebted countries ("contagion"), and (ii) a decrease in financial sounder countries' yields ("flight-to-quality").

Moreover, by looking at the current period, one important similarity with the pre-ESD crisis period emerges. Specifically, the global COVID pandemic crisis is putting pressure on sovereign debts and deficits, as the Great Financial Crisis of 2008-2009 also did in the pre-ESD crisis period. Therefore, we also aim at assessing whether a latent risk of sovereign bond market fragmentation is still present nowadays (given the current COVID global crisis).

Tackling fragmentation risk inside the euro area has been one of the major challenges for European authorities. In fact, a fragmented euro area sovereign bond market entails several economic and financial problems. First of all, the risk of a euro area breakup, and the subsequent risk that a euro-denominated asset gets devalued or redenominated into a new devalued legacy currency (De Santis, 2019), became sizeable and relevant for sovereign bond markets during the ESD crisis period. Moreover, within a fragmented bond market, both conventional and unconventional monetary policies' transmission is severely weakened and not effective (Grandi, 2019). Thirdly, euro area sovereign bond market fragmentation also implies several challenges in other markets. As documented by Macchiarelli and Koutroumpis (2016), existing literature proved how fragmentation in the euro area affects corporate bond markets (Zaghini, 2016), retail markets (Al-Eyd and Berkmen, 2013), interbank markets (Gabrieli and Labonne, 2018), and interest rate pass-through (Horvath, Kotlebova, and Siranova, 2018).

Given the crucial role played by fragmentation risk inside the euro area, this chapter aims at (i) modeling inter countries' sovereign bond markets' exposure, (ii) identifying the implications for the stability of the euro area, (iii) assessing European authorities' responses to both the ESD crisis and the current pandemic.

From a methodological standpoint this chapter reconciles Factor Augmented VAR ($FAVAR$) and Global VAR ($GVAR$) literature by employing (static) principal components (PC) as a way to retrieve information regarding cross-sectional dependence. Specifically, as in Elhorst, Gross, and Tereanu (2020), we confirm that in the presence of high cross-sectional correlation, the first PC captures the common shifts in the level of the entire cross-section of sovereign bond spreads. This factor can be interpreted as the common interdependence trend. Conversely, the second PC can be interpreted as an indicator of the fragility of the euro area sovereign bond market. Specifically, it measures how deep the difference between fragile and financially sounder economies is. We follow the intuition of Bernanke, Boivin, and Eliasz (2005), and retrieve the part of the second PC that is unspanned by the specific local market under analysis (i.e. *foreign* information). We express this component as the weighted average of foreign counterparts in the $GVAR$ fashion (Pesaran, Schuermann, and Weiner, 2004), but leaving the sign of the weights

unconstrained. Therefore, the analysis of the resulting transmission matrix W provides evidence of the presence of "flight-to-quality" and/or "contagion" effects in the euro area (Candelon and Luisi, 2020).

We apply this framework to the long-term yield spreads between the 10 year euro area government bonds and the German safe benchmark over three remarkable periods: the pre-ESD crisis, the post-ESD crisis as well as the recent COVID period. We highlight that signs of "contagion" and "flight-to-quality" effects were indeed already present in the pre-ESD crisis period, pointing out a latent fragmentation risk between peripheral European countries (Portugal, Italy, Ireland, Greece, and Spain) and the core ones (Austria, Belgium, Finland, France, and The Netherlands).

Soon after the ESD crisis, the different programs of public bonds purchases enhanced by the European Central Bank successfully reduced the fragmentation risk, leading all European spreads to shrink back slowly but steadily to the pre-ESD crisis levels.

The COVID pandemic has pushed again fragmentation risk at the forefront of the results. Despite the implementation of the recovery plan for Europe, "contagion" and "flight-to-quality" risks become evident once again, stressing the revival of fragmentation risk.

The chapter is sketched as follows: after a preliminary data analysis in Section 6.2, the methodology is presented in Section 6.3. The empirical analysis on the fragmentation risk is performed in Section 6.4, while Section 6.5 concludes.

6.2 Preliminary Data analysis

We start our analysis by examining the dynamics of the long-term yield differentials between euro area government bonds. The 10-Year ($10Y$) yields are extracted from the Refinitiv Eikon data set and are covering 11 countries (Austria, Belgium, Finland, France, Germany, Greece, Ireland, Italy, The Netherlands, Portugal, and Spain) over the period June 2006 - March 2021 and shown in Figure 6.1. Data are weekly.

What emerges from a first visual inspection is that long-term yields differentials between euro area government bonds and the German *safe* benchmark co-move following an unstable pattern over time.

In the following subsections, we give a closer look at this dynamics, distinguishing across the pre- (June 2^{nd}, 2006 - March 26^{th}, 2010), during- (April 2^{nd}, 2010 - July 27^{th}, 2012), and post- (August 1^{st}, 2012 -), ESD crisis periods. On the one hand, April 2010 represents the indicative watershed that certifies the beginning of the European sovereign debt crisis. In particular, after the publication of the Greek quarterly national accounts data that signaled a persistent recession for the Hellenic economy starting in 2007, credit rating agencies downgraded the Greek sovereign bonds to investment grade status in late April 2010, meanwhile, Prime Minister George Papandreou formally requested an international bailout for Greece. On the other hand, on July 26^{th}, 2012 the President of the European Central Bank (ECB) Mario Draghi pronounced the famous "Whatever it takes" speech, announcing the Outright Monetary Transactions ($OMTs$) to neutralize the risk of a euro breakup.[1]

[1]The verbatim of the speech by Mario Draghi is available at `https://www.ecb.europa.eu/press/key/date/2012/html/sp120726.en.html`

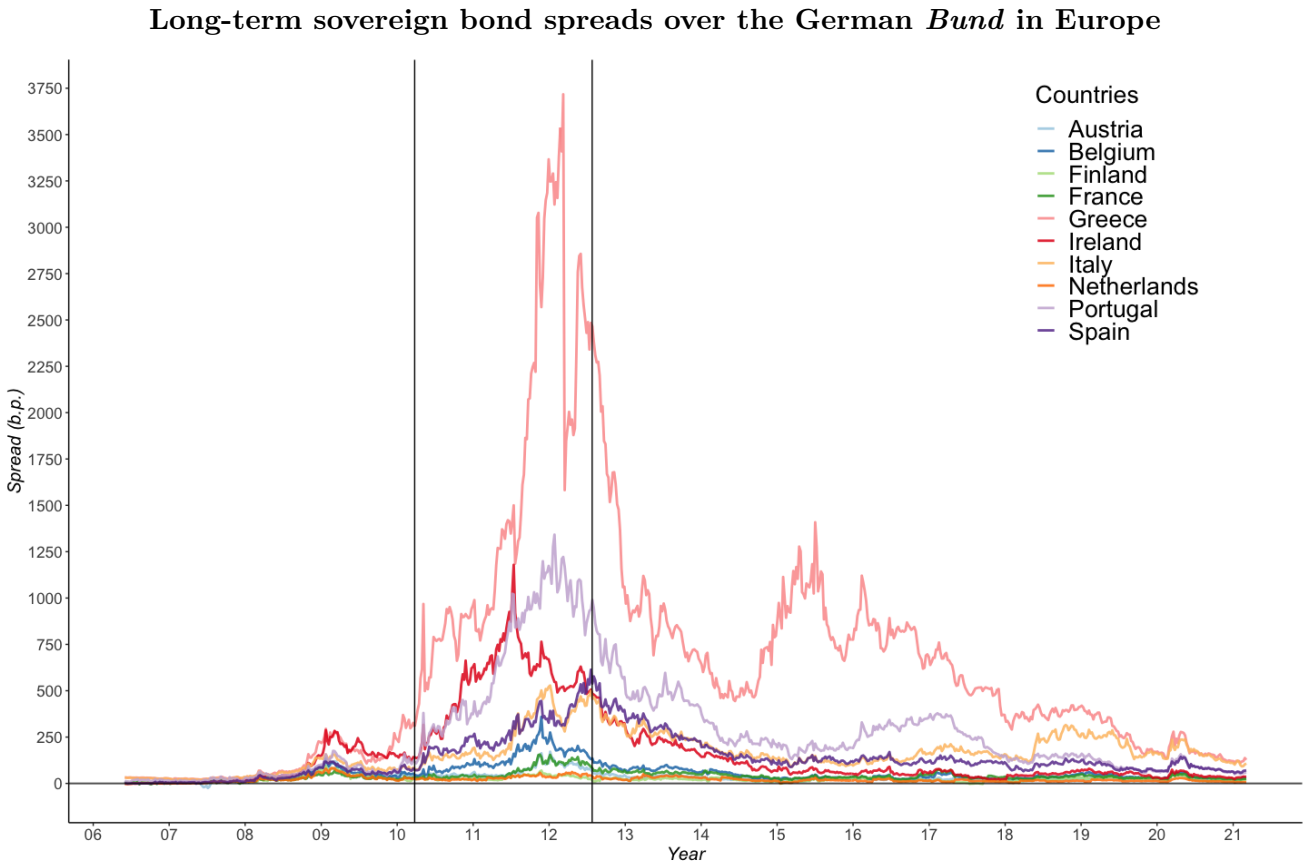


Figure 6.1: The figure reports the evolution of the government bond yield spread of 10 countries (Austria, Belgium, Finland, France, Greece, Ireland, Italy, the Netherlands, Portugal, and Spain) vis á vis of Germany over the period 2006 June - 2021 March. The *EDS* crisis beginning and end dates are indicated with vertical bars.

6.2.1 Pre-*ESD* crisis period

The dynamics of the spreads during the pre-*ESD* crisis period can be further decomposed into a convergence period and the subprime loans crisis period. From the introduction of the euro as a common currency, interest rate differentials co-moved synchronously at a very low level until the burst of the subprime loans crisis. As a reaction to the outburst of the Global Financial Crisis, sovereign spreads substantially increased. The dimension of such an increase firstly proved that investors never regarded different sovereign bonds inside the euro area as perfect substitutes (Favero and Missale, 2012). However, it is worth noticing that, as reported by Copelovitch, Frieden, and Walter (2016), during the Great Financial Crisis, European policymakers still regarded the European Monetary Union as an "unquestionable success" and a "rock of macroeconomic stability" helping Europe to weather the Global Financial Crisis.

6.2.2 The *ESD* crisis period

The *ESD* crisis outburst marked the beginning of the euro area fragmentation era. The idea that long-term government debt spreads follow a common dynamics in the euro area fades away: the Core (Germany, France, Belgium, Netherlands, and Finland) and Periphery

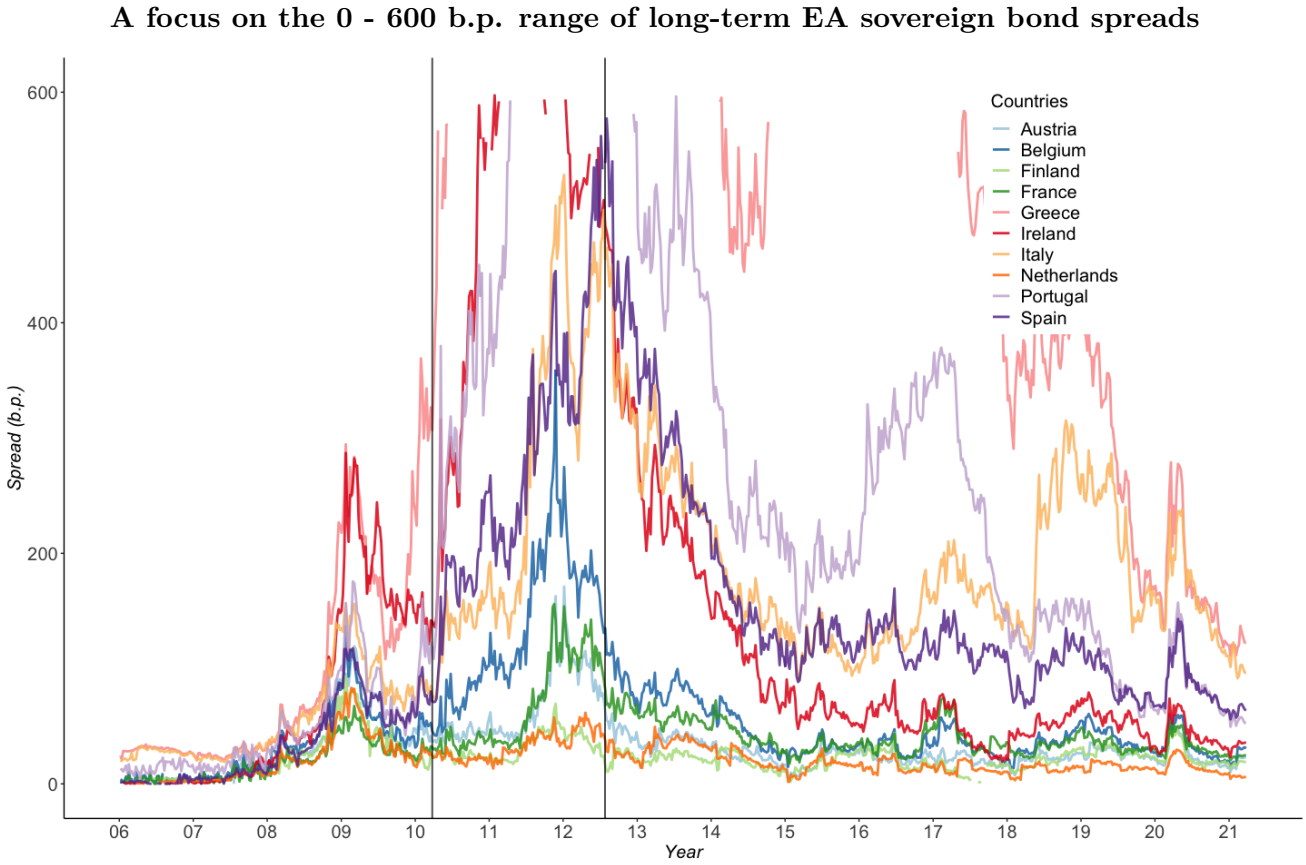


Figure 6.2: The figure reports the evolution of the government bond yield spread of 10 countries (Austria, Belgium, Finland, France, Greece, Ireland, Italy, the Netherlands, Portugal, and Spain) vis á vis of Germany over the period 2006 June - 2021 March. The EDS crisis beginning and end dates are indicated with vertical bars.

(Greece, Ireland, Italy, Portugal, and Spain) binary classification becomes predominant both in the political and academic debate when analyzing the economic and financial turmoil in $2010-2012$ (House, Proebsting, and Tesar, 2020; Corsetti, Kuester, Meier, and Muller, 2014).

During the fall of 2011, the $10Y$ differentials relative to the German *Bund* reached record heights: Italian $10Y$ sovereign yield spreads rising in a few weeks from 270 basis points in August to over 500 basis points in November. We can draw a similar picture for Ireland, Portugal and Spain. However, such outcomes were only partially connected to underlying fiscal fundamentals (De Grauwe and Ji, 2012), as they mainly reflected self-fulfilling prophecies and the existence of multiple *equilibria* for sovereign bonds' dynamics (Corsetti and Dedola, 2016; Cole and Kehoe, 2000). In early 2012, the risk of a euro area breakup became so relevant that the Euro redenomination risk, i.e., the risk that a euro asset is redenominated into a devalued legacy currency, significantly impacted sovereign yield spreads, with Italian and Spanish bonds being most negatively affected (De Santis, 2019).

Following Mario Draghi's, "Whatever it takes" speech, the ECB announced a new set of unconventional measures, the Outright Monetary Transactions ($OMTs$). This program explicitly aimed at relieving the financial fragmentation in the euro area and at neutralizing redenomination risk. As a fact, by September 2012, Irish, Italian, Portuguese and Spanish

differentials relative to the German $10Y$ yield had already decreased by an average of 200 basis points. Through the public purchasing programs (Securities Markets Programme), the very long-term refinancing operations ($VLTROs$), and the announcement of the $OMTs$, the ECB mitigated the fears in sovereign bond markets, reinforcing a slow but progressive return of spreads towards normal levels.

6.2.3 The post-ESD crisis period and the COVID pandemic.

The post-ESD crisis starts in August 2012, spanning the slow but steady return of spreads towards pre-crisis levels. In late 2012, ECB's actions addressed the difficulties in the transmission of monetary policy measures, the onset of a credit crunch, and the risk of deflation.[2]

As a result of these measures, the consolidated balance sheet of the Eurosystem has more than doubled between 2010 and 2020, starting from $2,002,210$ EUR million in 2010 and reaching $4,671,425$ EUR million in 2019.[3] Consequently, given the massive liquidity injections operated by the ECB during the 2012 - 2020 period, the progressive decline of the $10Y$ government bond differentials relative to the German *Bund* should not come as a surprise.

The COVID-19 pandemic outburst caused a sudden increase in sovereign bond spreads around March 2020 when the European spread of the virus caused the enhancement of the first traveling restrictions and lockdown measures across Europe. During this period, the European Commission, the European Parliament, and countries of the European Union have jointly been seeking an agreement to help repair the economic and social damages caused by the Coronavirus pandemic and boost the recovery at the European level. In December 2020, the agreement for a 1.8 trillion EUR recovery plan was reached[4] together with the return of generalized common and declining co-movement in the dynamics of $10Y$ government debt spreads. Among the novelties of the recovery plan, it is worth noticing that it will be issued by the European Union as a whole and that the associated debt will not be recorded in the countries' national accounts. It should result in reducing the risk premium and foster the link among European countries.

6.3 Empirical analysis

6.3.1 Modeling sovereign bond spreads in the euro area

GVAR Models (Pesaran, Schuermann, and Weiner, 2004) have been employed in the analysis of the interactions among European sovereign bonds spreads as a tool to unveil

[2]More precisely, in July 2013, the ECB began using forward guidance with the goal of strengthening the monetary policy transmission mechanism from short to long-term interest rates. In June 2014, the ECB further cut the main marginal lending facility rate to 0.4%, the refinancing operation rate to 0.15% and lowered the deposit facility rate to -0.10%. A series of the targeted longer-term refinancing operations ($TLTROs$) (June 2014, March 2016 and March 2019) have also been used to stimulate bank lending to the real economy. Starting in mid-2014, the ECB also launched the Asset Purchase Programme (APP), conducting net purchases of corporate sector ($CSPP$), public sector ($PSPP$), asset-backed securities ($ABSPP$) and covered bond ($CBPP3$).

[3]The Eurosystem consolidated balance sheet is published together with the Annual Accounts and is included in the Annual Report of the ECB. The data is available at `https://www.ecb.europa.eu/pub/annual/balance/html/index.en.html`

[4]see `https://ec.europa.eu/info/strategy/recovery-plan-europe_en`.

the tightness of financial interconnectedness and to identify the transmission channels of asymmetric shocks. Favero (2013) showed that the cross-correlation among sovereign bond spreads follows an unstable pattern over time, implying that a correct proximity measure should take into account the temporary shifts in investors' preferences when modeling interconnections in the euro area. Moreover, Favero and Missale (2012) showed how a shift in market sentiment following the emergence of a country's financial distress can propagate to relatively safer countries. For example, they estimate that a shock in August 2011 could trigger a contagion on the Italian sovereign yield of the size of 200 basis points.

Following Favero and Missale (2012), the local *GVAR* specification in the case of long-term sovereign bond spreads over the German *Bunds* ($Y_t^i - Y_t^{DE}$) can be expressed as follows:

$$(Y_t^i - Y_t^{DE}) = \beta_{i0} + \beta_{i1}(Y_{t-1}^i - Y_{t-1}^{DE}) + \beta_{i2}(Y_t^i - Y_t^{DE})^* + \beta_{i3}(Y_{t-1}^i - Y_{t-1}^{DE})^* + u_t^i \quad (6.1)$$

where $i = 1, \ldots, N$ is the specific country under analysis, β_{i0} the deterministic component, β_{i1} the coefficient associated with the country-specific lagged variable, and β_{i2} and β_{i3} the coefficients associated with the (contemporaneous and lagged) *Global Spreads*, $(Y_t^i - Y_t^{DE})^*$. *Global Spreads* are computed for each country as the weighted average of foreign counterparts' spreads over the German *Bund*, $\sum_{i \neq j} w_{ij}(Y_t^j - Y_t^{DE})$ for each $i \neq j$.

As recommended by Elhorst, Gross, and Tereanu (2020), we first perform the *CD*-test (Pesaran, 2021) on the cross-section of spreads to prove their strong co-movement along a global trend (see Section 6.4). Based on this result, we augment the *GVAR* specification in ((6.2)) as follows:

$$(Y_t^i - Y_t^{DE}) = \beta_{i0} + \beta_{i1}(Y_{t-1}^i - Y_{t-1}^{DE}) + \beta_{i2}(Y_t^i - Y_t^{DE})^* + \beta_{i3}(Y_{t-1}^i - Y_{t-1}^{DE})^* + \beta_{i4}\hat{F}_{1,t} + u_t^i \quad (6.2)$$

where $\hat{F}_{1,t}$ is the first Principal Component (PC) extracted from the cross-section of spreads.[5]

The common factor $\hat{F}_{1,t}$ can therefore be interpreted as a long-run stable attractor proxying the global interdependence among European countries' sovereign bonds. By augmenting our specification in this way, we make sure that the *star* variables in ((6.2)) and the associated coefficients deal with *intra-European* interconnection defined in a *GVAR* fashion.

The choice of the weights employed to build the *Global Spreads* is crucial. The *GVAR* specification demands the weights to summarize the *external* information through a weighted average of foreign counterparts. Existing literature has shown that misspecifying the channel of interconnection results in incorrect inferences (see, for example, Gross, 2019). Candelon and Luisi (2020) show that in the case of euro area sovereign bond spreads, an empirically valid channel of interaction should take into account the asymmetric transmission of shocks among countries that exhibit positive (negative) interconnections.

Therefore, it is natural to start the interconnection analysis from the second factor of the cross-section of sovereign spreads for several reasons. First, the second Principal Component $\hat{F}_{2,t}$ summarizes the cross-sectional information that is linearly independent to the first factor $\hat{F}_{1,t}$, therefore its inclusion in ((6.2)) would not be problematic from an

[5]Observable common variables are often also employed in the literature for this purpose (for example, Favero, 2013, augments the specification with the long-term corporate *Baa-Aaa* spread to account for global risk aversion). We refer to the A6.4 for a supplementary analysis of the correlations between the first PC and such observable variables.

econometric perspective. Second, by construction $\hat{F}_{2,t}$ is designed to explain most of the cross-sectional variance left unexplained by $\hat{F}_{1,t}$, capturing, therefore, the interconnection among spreads as we aim to. Third, the loadings of the factors are not constrained to be nonnegative, providing guidance for not only the size but also the sign of the weight of each local yield contribution to the cross-sectional variance explained. Four, it allows us to follow the intuition in Bernanke, Boivin, and Eliasz (2005) and extract from the Principal Component the *foreign* information, as requested by the *GVAR* framework, where the *Global* variable is a weighted average of foreign counterparts. Finally, the factor summarizing external information as a (weakly) exogenous variable can be included in a typical *VARX** fashion (see, for example, Balabanova and Brüggemann, 2017).

For all these reasons, we consider:

$$\hat{F}_{2,it} = \sum_{i \neq j} w_{ij}(Y_t^j - Y_{t-1}^{DE}) + v_{it} \tag{6.3}$$

for every $i = 1, \ldots, N$ with $\sum_{i \neq j} |w_{ij}| = 1$. $\hat{F}_{2,it}$ is part of the second Principal Component that is unspanned by the local economy i taken into account (see A6.1). Usually in the *GVAR* literature, the weights are constrained to be nonnegative. However, the normalization we propose allows keeping the original properties of the size of the weights (Pesaran, Schuermann, and Weiner, 2004), while considering also that sovereign bond spreads' interconnections can be positive and negative. Relaxing the nonnegativity assumption becomes particularly relevant when mechanisms as "contagion" (see Metiu, 2012; Candelon and Tokpavi, 2016) or "flight-to-quality" (Monfort and Renne, 2013) are present. Specifically, during distressed periods, investors prefer safer and more liquid assets. This would result in an outflow of money from the peripheral to the Core countries, causing rising yields for the highly indebted euro area members ("contagion") while financially sounder economies could benefit from lower rates ("flight-to-quality"). In our framework, this possibility is not ruled out thanks to the unrestricted sign of the weights.

Once the weights are retrieved by estimating ((6.3)) with OLS, we can build the foreign weighted averages and estimate the local *VARX** systems as specified in ((6.2)).

Armed with the local coefficients, we can proceed to solve the model in order to retrieve the global reduced form *VAR* representation (see A6.2) useful for Generalized Impulse Response analysis.

6.3.2 A closer look to the first and second PCs of the cross section of euro area spreads.

In our model, $\hat{F}_{1,t}$ refers to the first principal component (PC_1) of the cross section of sovereign bond spreads. As we can see in Figure 6.3, its loadings are all roughly of the same dimension and sign across countries. Consequently, the first principal component can also be seen as an equally weighted portfolio of all sovereign bonds (up to a scaling factor). To correctly interpret this feature, we refer to the *Level* factor in the term structure literature for a meaningful comparison. Specifically, as in the term structure literature (see, for example, Joslin, Singleton, and Zhu, 2011), the *Level* factor (that exhibits similar weights across maturities) captures shifts in the level of the entire yield curve (across all maturities), in our case, $\hat{F}_{1,t}$ captures shifts in the level of the entire cross-section of long-term spreads, what we define *interdependence* trend.

Conversely, the loadings of the second principal component (PC_2) are heterogeneous both in sign and dimension. Specifically, across all the time periods considered, they show opposite signs between core and peripheral countries. The second principal component of the term structure of yields (i.e., the *Slope* factor) shows similar characteristics for short and long-term maturities. As the *Slope* factor measures how steep the yield curve is, so in our model the second principal component measures how deep the difference between fragile and financially sounder economies is. Therefore, we name this factor the *intra-European fragility* factor.

The relative share of each country in its group is also important for inference. During the pre-ESD crisis period, Italy shows the same sign as low-yield countries. However, its relative weight is only $0.07/1.3 = 5.4\%$ while, for example, Finland represents $0.458/1.3 = 35.2\%$. The relative weights translate into the Italian spread being associated with a peripheral behavior (i.e., subject to "contagion"). Lastly, the relative share of low and high-yield countries in the PC_2 is also easily interpretable. If we look at the post-ESD crisis period, for example, the sum of weights of low-yield countries increases to 1.4 while the share of high-yield countries decreases to 1.168. This translates into a more stable and integrated euro area (no asymmetric exposure to shocks from peripheral countries, even if the dimension remains different).[6]

Loading of the first two Principal Components of the cross-section of spreads

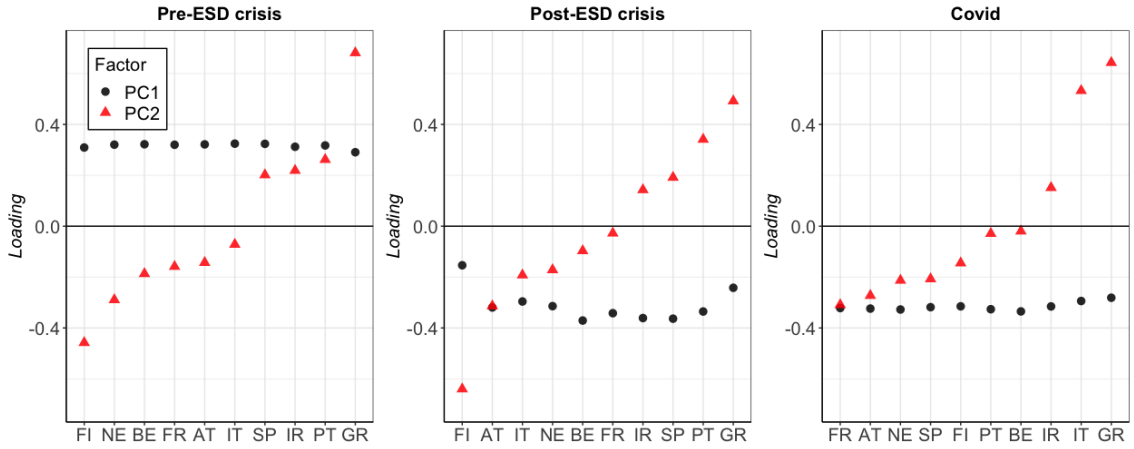


Figure 6.3: The figure reports the value of the cross-sectional loadings of the first two Principal Components of the government bond yield spread of 10 countries (Austria, Belgium, Finland, France, Greece, Ireland, Italy, the Netherlands, Portugal, and Spain).

6.4 Results

6.4.1 Preliminary results

Following Favero and Missale (2012), and given the data analysis outlined in Section 6.2, we consider two separate periods: pre-ESD crisis (June 2006 to March 2010) and post-ESD crisis (from August 2012 to February 2019). Moreover, the recent COVID period is separately analyzed to further study the last part of the sample, given that,

[6]These facts will be also confirmed by the analysis of the Generalized Impulse Responses in Section 6.4

during a period of co-movement at low levels, sovereign bond spreads are exposed to the global pandemic.

The sample split offers several advantages. First of all, by considering homogeneous periods in terms of cross-correlation of the spreads, we can assess the changes in interconnectedness in those different phases (Forbes and Rigobon, 2002), while accounting for the observed shifts in market sentiment (Favero, 2013). Furthermore, using a narrative approach, the effect of unconventional monetary policies in fostering financial integration can be assessed as in Favero and Missale (2016).

The first result reported in Table 6.2 regards the *CD*-test to assess the cross-sectional dependence of the sovereign spreads under analysis. The *CD*-test (Pesaran, 2021) is defined as:

$$CD = \sqrt{\frac{2T}{N(N-1)}} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \hat{\rho}_{ij}, \tag{6.4}$$

with $\hat{\rho}_{ij}$ being the sample correlation coefficient between local variables of county i and country j. In Bailey, Holly, and Pesaran (2016), we can find the order property of the average correlation coefficient:

$$\bar{\rho}_N = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \rho_{ij} = O(N^{2\alpha-2}), \tag{6.5}$$

where $\alpha \in [0, 1]$. As recommended in Elhorst, Gross, and Tereanu (2020), we first control that the null of weak cross-sectional dependence ($\alpha < 0.5$) cannot be rejected in our sub-samples through the *CD*-test ($\hat{\bar{\rho}}$ and CD-statistic's columns in Table 6.2). In fact, weak cross-sectional dependence cannot be rejected given the high value of $\hat{\bar{\rho}}$.

Secondly, we assess the degree of cross-sectional correlation through the two-step procedure as in Bailey, Holly, and Pesaran (2016).[7] In Table 6.2 we show that the cross-sectional correlation in the sub-samples under analysis is really strong. Specifically, the estimated α is larger than 3/4, clearly pointing to the presence of common components at the basis of the behavior of sovereign spreads. To correctly capture the interconnections among spreads, it is therefore important to *defactorize* the data. Following Vega and Elhorst (2016), we opt for a simultaneous approach as specified in ((6.2)).

To further document that *defactorizing* the data is important (and that no interconnection among units is considered at this stage), we show in Table 6.4 the loadings of the first Principal Component. As expected in such a case, the loadings are evenly distributed among the different sovereign spreads under analysis. Hence, it supports the evidence of a persistent interdependence trend among European sovereign yield spreads (as illustrated in sub-section 6.3.1).

6.4.2 Interconnection matrices and their signs

We now proceed to the analysis of the estimated weights. As described in Subsection 6.3.1, we computed the weights by estimating ((6.3)) via OLS. The weights describe the contribution of each $j \neq i$ to the space of the residual unexplained cross-sectional variance that is unspanned by country i's sovereign yields, as in Bernanke, Boivin, and Eliasz (2005).

[7] In the empirical application, we estimate α as in Bailey, Kapetanios, and Pesaran (2016)

Table 6.2: Notes: This table reports the estimated average sample correlation between local country i and j, $\hat{\bar{\rho}}$, the cross-sectional test statistics, $CD - statistics$, as well as $\hat{\alpha}$ and its standard error, $\hat{\sigma}_{\hat{\alpha}}$

	$\hat{\bar{\rho}}$	CD-statistic	$\hat{\alpha}$	$\hat{\sigma}_{\hat{\alpha}}$
Pre-crisis	0.964	91.415	0.7559	0.0213
Post-crisis	0.945	133.355	0.8463	0.0258
COVID-19	0.918	63.409	0.9013	0.0499

Table 6.4: Notes: This table reports the factor loadings of the first static Principal Component over the three periods of investigation and the portion of cross-sectional variance explained.

	Pre-ESD	Post-ESD	COVID
Austria	0.321	-0.319	-0.324
Belgium	0.322	-0.371	-0.335
Finland	0.309	-0.154	-0.315
France	0.320	-0.342	-0.322
Ireland	0.312	-0.361	-0.315
Italy	0.324	-0.296	-0.294
The Netherlands	0.320	-0.314	-0.327
Portugal	0.317	-0.335	-0.326
Spain	0.324	-0.363	-0.318
Greece	0.291	-0.242	-0.281
Variance explained	0.922	0.676	0.851

In Table 6.5, we report the weights, row normalized. Therefore, each row i features the tightness and the sign of the interconnection among country i and the foreign counterparts j. By looking at the values, we can easily notice the presence of sign heterogeneity, some weights being positive, while others negative. This feature corresponds to the first signal of the presence of asymmetric interconnection. Specifically, already during the pre-ESD crisis period it was clear the distinction that emerged afterward in the academic and public debate between Core and peripheral countries. The sovereign yields of Spain, Italy, Portugal, and Greece feature an opposite sign for almost all specifications compared with Austria, Belgium, Finland, France, and The Netherlands. Such a result portrays the "flight-to-quality" phenomenon from peripheral to Core European countries already before the outburst of the ESD crisis[8].

During the post-ESD crisis period, while Portugal and Greece continue to exhibit the above-mentioned features, we can see mixed signs for Ireland and Italy. This result points out the convergence path started by European countries after the introduction of unconventional monetary policies inside the euro area.

[8]This result over the pre-ESD crisis sample is coherent with the analysis of Candelon and Luisi (2020) during the period 2001 - 2009. The two periods are indeed similar in terms of cross-country sovereign spread correlations.

Table 6.5: Notes: This table reports the weights w_{ij} extracted as indicated in (A6.3).

a) Pre-*ESD* crisis period

	AU	BE	FN	FR	IR	IT	NL	PT	SP	GR
AU	·	-0.067	-0.258	-0.138	0.035	-0.071	-0.178	0.066	0.089	0.098
BE	-0.037	·	-0.263	-0.158	0.040	-0.088	-0.169	0.060	0.084	0.101
FN	-0.023	-0.083	·	-0.161	0.037	-0.164	-0.233	0.059	0.130	0.110
FR	-0.044	-0.107	-0.275	·	0.036	-0.093	-0.172	0.058	0.106	0.111
IR	-0.056	-0.138	-0.261	-0.089	·	0.002	-0.168	0.085	0.109	0.092
IT	-0.050	-0.086	-0.297	-0.164	0.036	·	-0.132	0.053	0.093	0.089
NL	-0.046	-0.075	-0.298	-0.161	0.039	-0.111	·	0.050	0.104	0.115
PT	-0.086	-0.101	-0.344	-0.071	0.031	0.045	-0.153	·	0.053	0.116
SP	-0.074	-0.078	-0.290	-0.122	0.038	0.006	-0.196	0.096	·	0.099
GR	-0.072	-0.097	-0.205	-0.080	0.022	0.086	-0.224	0.127	0.088	·

b) Post-*ESD* crisis period: W matrix

	AU	BE	FN	FR	IR	IT	NL	PT	SP	GR
AU	·	0.026	-0.140	-0.377	0.058	0.063	-0.124	0.040	-0.126	0.047
BE	-0.177	·	-0.107	-0.256	0.068	0.061	-0.210	-0.002	-0.067	0.050
FN	-0.213	0.027	·	-0.317	0.063	0.060	-0.147	0.025	-0.102	0.047
FR	-0.282	0.111	-0.156	·	0.040	0.038	0.007	0.079	-0.232	0.056
IR	-0.165	-0.050	-0.101	-0.174	·	0.066	-0.331	-0.042	-0.013	0.058
IT	-0.192	-0.078	-0.051	0.062	0.068	·	-0.338	-0.093	0.064	0.054
NL	-0.176	0.028	-0.124	-0.344	0.064	0.065	·	0.033	-0.117	0.049
PT	-0.181	-0.002	-0.110	-0.262	0.068	0.061	-0.199	·	-0.067	0.050
SP	-0.143	0.046	-0.142	-0.481	0.042	0.059	-0.015	-0.020	·	0.052
GR	-0.124	-0.090	-0.032	-0.082	0.092	0.069	-0.425	-0.046	0.040	·

c) COVID period: W matrix

	AU	BE	FN	FR	IR	IT	NL	PT	SP	GR
AU	·	-0.091	-0.683	0.052	0.0002	-0.015	-0.088	0.023	0.031	0.017
BE	-0.215	·	-0.509	0.098	-0.035	0.00003	-0.085	0.010	0.033	0.016
FN	-0.227	0.115	·	-0.092	0.048	-0.065	-0.394	0.013	0.023	0.022
FR	-0.211	-0.063	-0.511	·	-0.057	0.014	-0.061	0.019	0.047	0.017
IR	-0.137	-0.227	-0.360	0.200	·	0.017	0.030	0.007	-0.008	0.014
IT	-0.162	-0.026	-0.478	0.108	-0.026	·	-0.137	0.007	0.044	0.012
NL	-0.172	-0.070	-0.557	0.107	-0.025	-0.006	·	0.013	0.035	0.016
PT	-0.176	-0.209	-0.317	0.169	-0.039	0.025	0.033	·	0.018	0.014
SP	-0.158	-0.180	-0.390	0.187	-0.026	0.017	0.021	0.007	·	0.015
GR	-0.124	-0.219	-0.272	0.209	-0.050	0.022	0.049	0.010	0.044	·

By closely looking at the COVID period of our sample, we can clearly see that the same signs of fragility are back in the sovereign bond market. As a matter of fact, the blocks of countries exhibiting opposite signs are again Core and Peripherals, with the

exception of Belgium, showing some similarities with peripheral countries.

6.4.3 Impulse response analysis

After the analysis of the transmission matrix W, we further evaluate the properties of the $GVAR$ model by studying the impulse response functions. This analysis provides useful insights when we study multivariate dynamic systems. In particular, the impulse response functions map the responses of the system to the effect of shocking one of the variables, at different horizons.

In our GVAR, we have three channels of transmissions of a shock from country i to j: a) via the contemporaneous factor $\hat{F}_{2,t}$, b) its lag $\hat{F}_{2,t-1}$ and c) via the covariance matrix of the errors. Therefore, it is difficult to have a clear identification strategy both when choosing the relevant restrictions in the covariance matrix (Bernanke, 1986; Sims, 1986; Blanchard and Quah, 1989) or when deciding the order of countries and variables in the Orthogonalized Impulse Responses (OIR) (Sims, 1980) set up. The Generalized Impulse Response ($GIRF$) (Pesaran and Shin, 1998; Pesaran, Shin, and Smith, 1999) is the standard tool to overcome these limitations in a GVAR framework. $GIRFs$ are invariant to the ordering of the variables and the countries in the $GVAR$ and does not require any restrictions on the covariance matrix of the residuals. In fact, the $GIRF$ considers system shocks and it consists of inducing a one-period shock only on one variable, e.g., the $10Y$ spread of country i, and then, ruling out the historically observed distribution of the errors, predicting the effect of such shock on the system at different horizons.

In the following subsections, we present the generalized impulse response functions in the pre- and post-ESD crisis with a supplementary focus on the COVID pandemic. We build the confidence interval using $10,000$ bootstrap replications (see A6.3 for more details), indicating with a red solid line the median response and with black dotted lines the 16^{th} and 84^{th} confidence bounds.

6.4.4 Pre-ESD crisis period

Figure 6.4 reports the generalized impulse responses to a 200 basis points (b.p.) increase of the Spanish $10Y$ differentials. Spain was selected as it is a large peripheral country, which has been quite affected during the ESD crisis period but also by the COVID pandemic.[9]

Our model predicts a fragmentation of the euro area into Core and Periphery even before the beginning of the European sovereign debt crisis (our sub-sample precedes the ESD crisis period). Italy, Ireland, Portugal, and Greece respond in a similar manner: a 200 b.p. increase in Spanish differentials is associated with increasing differentials in the other peripheral countries.

Besides the sign, the size of such a response is also relevant: the estimated increase is well beyond 60 b.p., (150 b.p. for Greece) with the only exception being Portugal. However, the response of Portuguese differential is not statistically significant up to two weeks horizon, but then the predicted increase of the spread becomes significant and particularly persistent.

[9]Italy for example has been less affected in term of GDP by the recent pandemic. We refer to the Supplementary Material for the $GIRF$s to a 200 b.p. increase in the spreads of the remaining peripheral countries.

Conversely, with the exception of Belgium, the 200 b.p. increase of Spanish differentials is associated with a decrease in core economies' spreads: our model predicts the "flight-to-quality" mechanism that will be typically studied in the context of the European Sovereign Debt crisis. A second important aspect that will be a common feature of our results is the humped shape of the $GIRF$. In fact, while the first transmission of the shock occurs at the level of the covariance matrix of the $GVAR$, then subsequent effects and the persistence of the impact are mostly due to the *foreign* factor and its lag. The case of Belgium is exemplary of this: despite a first significant increase of its differential with respect to Germany (a stylized fact also underlined by Candelon and Luisi, 2020), the $10Y$ spread's response becomes insignificant, dropping in the mild negative territory, aligning Belgium to the other core economies.

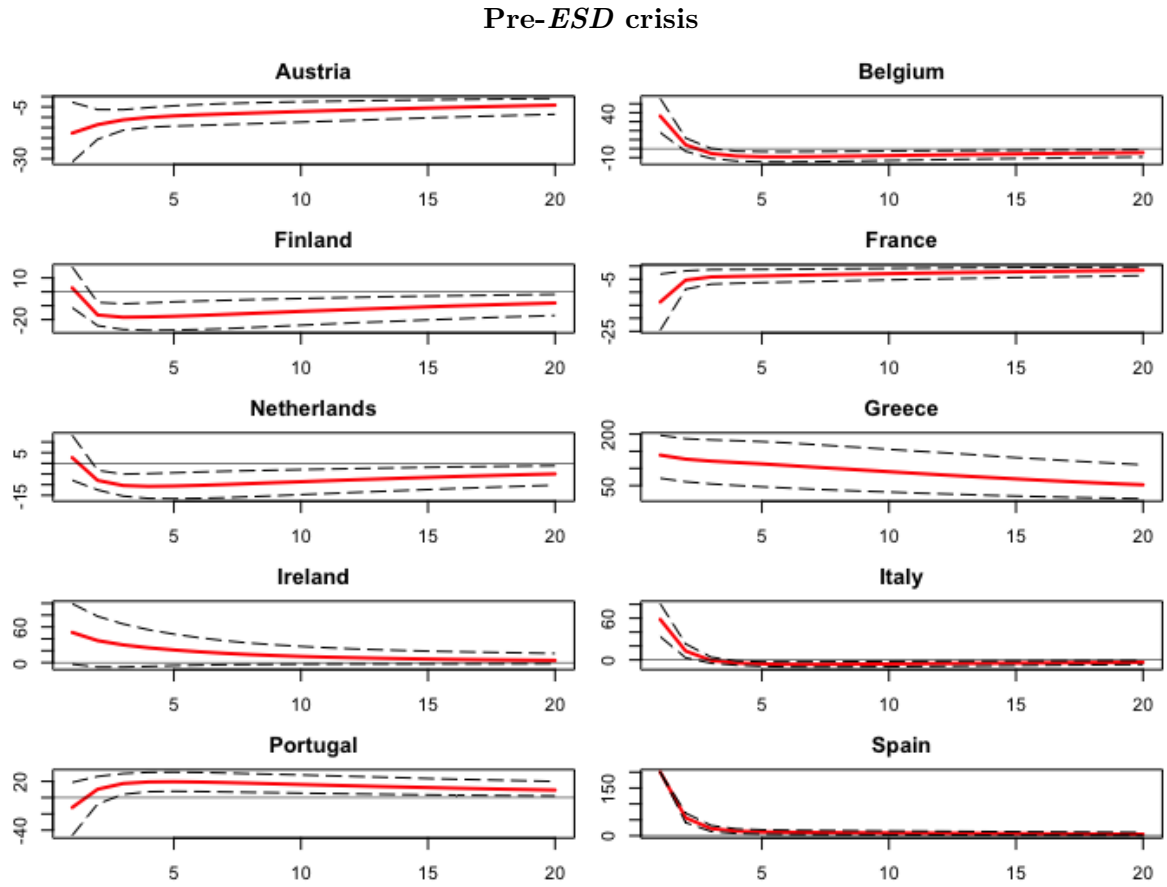


Figure 6.4: The figure reports the Generalized Impulse Response Functions to a 200 b.p. shock to the $10Y$ Spanish Spread in the pre-crisis period. $[16^{th}, 84^{th}]$ confidence intervals are calculated using $10,000$ bootstrap replications as described in the Appendix.

6.4.5 Post-ESD crisis period

Figure 6.5 reports the generalized impulse responses to a 200 basis points (b.p.) increase of the Spanish spread. Compared to the pre-ESD crisis period, some differences are worth to be underlined. The increase of the Spanish differentials is associated with a generalized statistically significant increase in the spreads of both peripheral and core economies: there are no signals of a "flight-to-quality" mechanism. Nevertheless, there

is still a marked difference in the size of the responses: while the response of the core economies is particularly contained in size, the effect of the same increase on the other peripheral country remains important (e.g., over 150 b.p. for Greece) and persistent.

Figure 6.5 depicts a picture where, despite the perception of the risk of a euro area breakup (no sign heterogeneity in the responses) seems to vanish, markets still acknowledge the financial fragility of the Periphery relative to the Core (size heterogeneity in the responses).

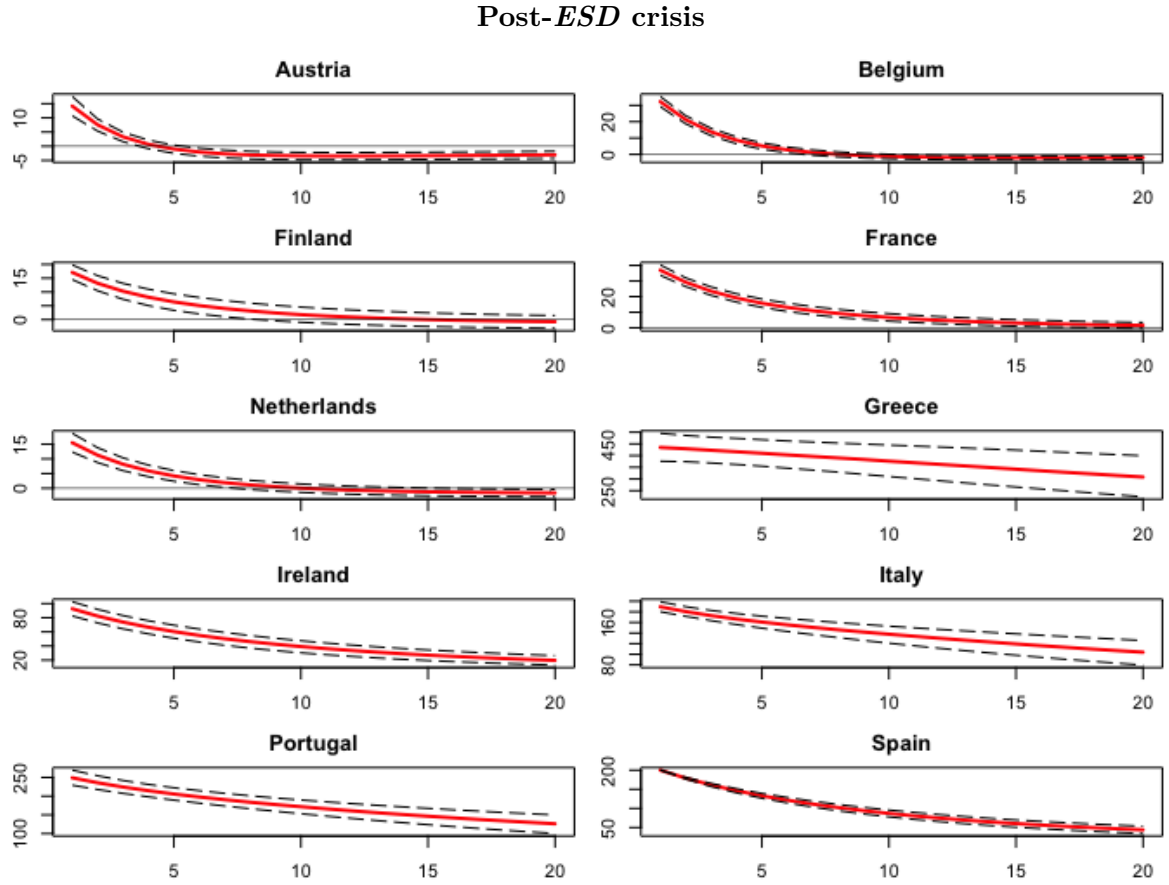


Figure 6.5: The figure reports the Generalized Impulse Response Functions to a 200 b.p. shock to the $10Y$ Spanish Spread in the post-crisis period. $[16^{th}, 84^{th}]$ confidence intervals are calculated using $10,000$ bootstrap replications as described in the Appendix.

6.4.6 The COVID pandemic period 2019 March - 2021 March

As in the previous analyses of the $GIRF$s, we still consider a 200 b.p. increase of the Spanish spreads also in Figure 6.6.

A question that naturally arises from the analysis of Figure 6.6 is whether the convergence phase that followed the post-crisis period is over.

In fact, while the size of the responses of the peripheral economies remains somehow unchanged compared to the post-ESD crisis period, Figure 6.6 depicts a completely different picture compared to Figure 6.5 with respect the responses of the core economies: the 200 b.p. increase of the $10Y$ Spanish differentials is now associated with a generalized and statistically significant decrease in the $10Y$ differentials in core economies. This result

highlights the revival of the "flight-to-quality" effect, leading to an increase in the *potential* risk of fragmentation of the euro area.

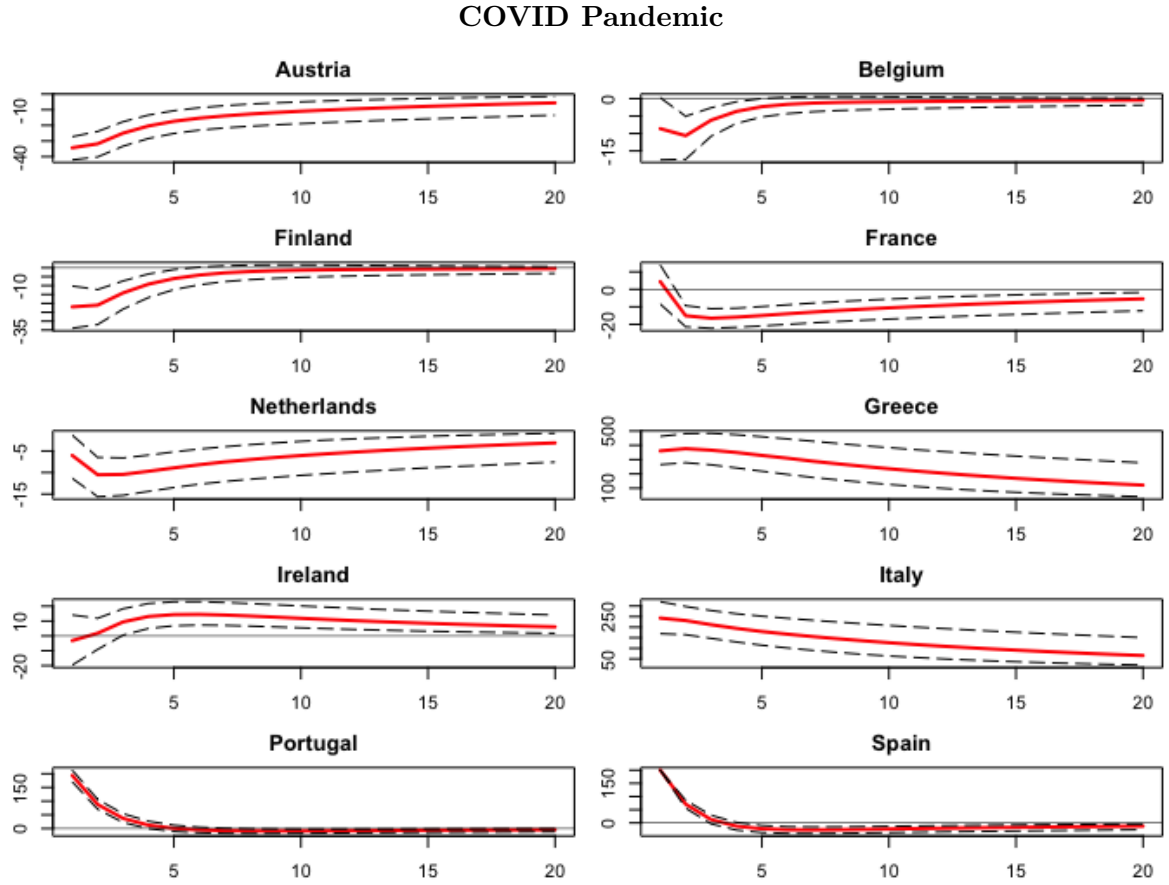


Figure 6.6: The figure reports the Generalized Impulse Response Functions to a 200 b.p. shock to the $10Y$ Spanish Spread in the COVID period. $[16^{th}, 84^{th}]$ confidence intervals are calculated using $10,000$ bootstrap replications as described in A6.3.

6.4.7 Heatmap: a closer look at the interconnections underlying the $GIRF$s

To summarize the message conveyed by $GIRF$s, we have reported in Figure 6.7 the heatmap defined by the average correlations of the different $GIRF$ after a 200 basis points (b.p.) shock to the Greek, Irish, Italian, Portuguese and Spanish $10Y$ differentials.

We hence focus on the signs of the responses associated with a shock in a peripheral country. To this goal, the heatmap is built on the correlation between the different $GIRF$s obtained after a shock in the spread of a peripheral country for a given forecasting horizon. In our case, we consider a 30 weeks forecasting horizon and the three selected periods: pre-ESD crisis, post-ESD crisis, and the COVID pandemic.

The fragmentation between Core and Periphery is evident in the pre-ESD crisis period. Ireland, Italy, Greece, Spain, and Portugal are strongly and positively correlated together but they are also negatively (or weakly) correlated to the Core European countries. As the sample ends before the ignition of the ESD crisis, this finding confirms the good predictive power of our $GVAR$ model with positive and negative weights in the transmission matrix

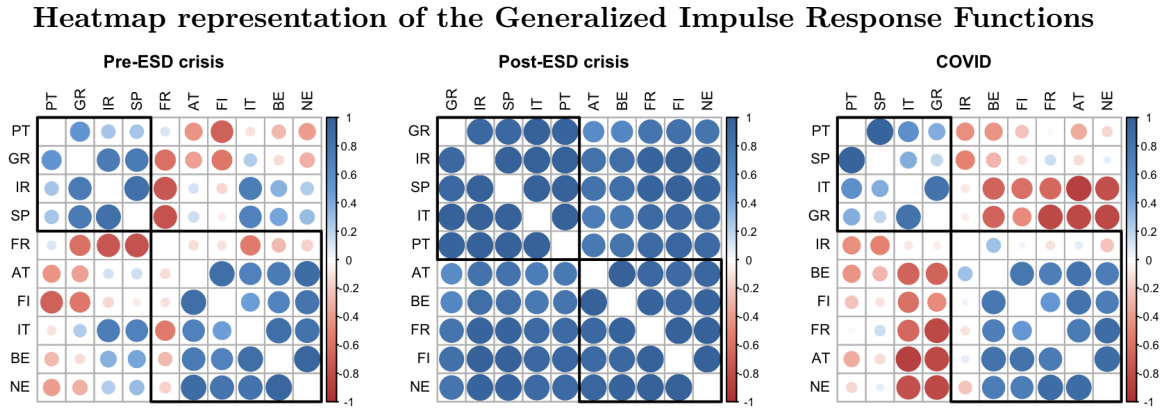


Figure 6.7: The figure displays the average heatmap representation of the Generalized Impulse Response Functions to a 200 b.p. shock to the $10Y$ Greek, Spanish, Portuguese, and Italian Spreads in the pre-crisis, post-crisis, and COVID pandemic period with 30 weeks forecasting horizon. We indicate in blue and red, the positive or negative signs of the correlation between the two countries, respectively. The shade of the color and the diameter of the circle represents the intensity of the correlation: the darker (the greater) the circle is, the stronger the correlation is. The rectangles identify the clusters defined by the dissimilarity measure $1 - \rho_{i,j}$, with $\rho_{i,j}$ being the pairwise average correlation coefficient between the $GIRF$ of country i and j to a 200 b.p. shock to Greece, Ireland, Italy, Portugal, and Spain.

W. Two countries appear to have a specific role. On the one hand, France is weakly negatively correlated to both peripheral and other Core countries. On the other hand, Italy is the important node that constitutes the junction between both groups: it is positively correlated to Ireland, Greece, and Spain but also positively correlated to Austria, Finland, Belgium, and the Netherlands. This strategic positioning reveals the key role played in the fragmentation risk by this country.

The post-ESD crisis heatmap shows that all the measures enhanced by the ECB have successfully neutralized the devaluation risk and created tighter links between European sovereign bonds. Furthermore, the signs of the interconnections are all strongly positive: this further underlines how, in this period, the euro area sovereign debts behaved as an integrated market. The COVID period heatmap surprisingly looks similar to the pre-ESD crisis one. The degree of similarity amounts to 51.71%, which is well above 0 and quite high when compared with the post-ESD crisis period (11.70%).[10] Both groups, Core and Periphery, are once again fragmented. Only three slight differences with the pre-ESD crisis can be noticed: first, while Italy joins the group of peripheral countries, Ireland becomes more positively (or weakly) correlated to the core economies, hence leaving the periphery. Second, France is reintegrating its positioning in the European Core. The similarities of the pre-ESD crisis and the COVID periods constitute therefore a clear signal of a resurgence of fragmentation risk in the European Monetary Union.

[10]We compute the degree of similarity between the heatmaps by the percentage average L1 norm of the difference between the correlation matrices underlying the heatmap representations in the pre-, post-ESD crisis and COVID periods.

6.4.8 Learning from the past: policy implications

The results of the empirical analysis point out at a recent rise of fragmentation risk inside the euro area, defined as heterogeneous exposure of local sovereign bonds to foreign counterparts in the monetary union. Furthermore, the analysis suggests that system shocks generate opposite patterns of responses between core and peripheral countries, as before the pre-ESD crisis outburst.

It is also interesting to notice two characteristics of the first principal component of the cross-section of sovereign spreads (a proxy for the interdependence trend). While in Table 6.4 we can see that the cross-sectional variance explained by this component raised from 0.67 (post-ESD crisis period) to 0.85 (COVID period), it is worth mentioning that it strongly correlates with the "flight-to-liquidity" measure of Monfort and Renne (2013) during the recent COVID period (see A6.4 for more details), pointing at a rise of capital flights to safe heavens, coherently with Cheung, Steinkamp, and Westermann (2020).

The similarities between the pre-ESD crisis and the COVID periods in terms of fragmentation risk, a resurgence of capital flights to safe heavens, in the context of weakened monetary policy transmission (Corsetti, Duarte, and Mann, 2020), together with an uncertain global outlook (International Monetary Fund, 2021), demand for adequate policy responses to avoid the triggering of a new sovereign debt crisis. The comparison between the pre-ESD crisis period and the current situation allows us to (i) delineate the successful characteristics of the Outright Monetary Transactions (and its eventual criticism), but most importantly (ii) to find correspondence between the $OMTs$ and current measures, allowing for an early evaluation.

In the context of self-fulfilling liquidity crises, the European Central Bank responded successfully with the Outright Monetary Transactions announcement. If we compare the OMTs with the previously implemented Security Market Program, the Outright Monetary Transactions brought the following novelties (i) no *ex-ante* limit to the purchases by the European Central Bank, (ii) no seniority status of OMT holdings, and (iii) commitment to fiscal reforms in the context of the European Financial Stability Facility/European Stability Mechanism ($EFSF/ESM$).[11]

However, as stressed in Favero and Missale (2016), the OMTs still exhibit some issues. Specifically, to activate the program, not only *ex-ante* eligibility criteria have to be met in terms of sustainable debt, deficit, and sound financial system, but must also *ex-post* additional reforms have to be put into action. If a country under liquidity pressure already satisfies the eligibility criteria, the requirement of additional fiscal adjustments would not only send a wrong signal to the market but also carry what the authors define as a *stigma* implied by the imposition of fiscal adjustments.

If we now look at the current public rescue package, we can outline some critical features. The first one regards seniority. One of the reasons the OMTs managed to tackle the panic on the financial markets is that the European Central Bank holdings no longer benefit from seniority status. Multilateral loans in the context of the current public rescue package are expected to have a senior stance in case of insolvency. As documented in Steinkamp and Westermann (2014), multilateral loans with senior status do relax the budget constraints of the countries, but do not necessarily have an impact on their effective interest rates.

[11] Another difference with the Security Market Program was the focus on maturities between one and three years.

Secondly, the whole rescue package has a fixed volume while OMTs had no *ex-ante* limit to the purchases. Finally, the Next Generation EU Recovery Plan suffers from the same *ex-post* conditionality that OMTs holdings had. In fact, partial disbursement of funds will be possible if the intermediate objectives are not achieved. The *ex-post* monitoring and the subsequent possibility of only partial disbursement of funds could cause important financial problems in terms of signaling.

Seniority status of the funds, *ex-ante* fixed volume, and *ex-post* conditionality of funds' disbursements appear to be the bottlenecks of the recent policy responses to ensure that the spreads of less financially sound economies will be more resilient to self-fulfilling liquidity crises.

6.5 Conclusion

This chapter sheds new light on the fragmentation risk in the European Monetary Union. A novel methodology reconciling Factor Augmented and Global VAR models allows disentangling of the interdependence from the "contagion" and the "flight-to-quality" effect. It is implemented for yield spreads between 10-year euro area government bonds and the German safe benchmark over three remarkable periods: the pre-ESD crisis, the post-ESD crisis as well as the recent COVID period. We show that signs of "contagion" and "flight-to-quality" effects were already present in the pre-ESD crisis period, pointing out the latent fragmentation risk between peripheral European countries (Portugal, Italy, Ireland, Greece, and Spain) and the core ones (Austria, Belgium, Finland, France, and The Netherlands). The different measures enhanced by the European Central Banks have limited this risk, leading all European spreads to shrink homogeneously in the post-ESD crisis period.

The COVID pandemic has reintroduced the fragmentation risk at the forefront of the European sovereign bond market for peripheral countries: similarly to the pre-ESD crisis period, an asymmetric local shock could transmit heterogeneously across countries. This is a new challenge for European policymakers. While the ECB's actions to maintain the stability of prices and the European Monetary Union under control remain somehow unchanged from the OMT announcement, the Next Generation EU Recovery Plan, leading to the issuance of European mutualized debt, is the first strong political step towards a deeper fiscal and financial integration in the European Union. The features of such rescue package (i.e., its seniority status, its *ex-ante* fixed size, the *ex-post* conditionality and public spending monitoring clauses), however, pose also some limits for tackling fragmentation risk in the euro area. If the long-term impact of such a package on the interest rates of the peripheral countries turns out not to be effective, fragmentation risk and the heterogeneous transmission of common shocks will remain undisputed obstacles for financial integration in the euro area.

APPENDIX

A6.1 How to retrieve the country-specific weights

Following the intuition of Bernanke, Boivin, and Eliasz (2005), we employ the following methodology to extract foreign information from the second Principal Component of the cross-section of spreads:

1. Extract the second Principal Component ($\hat{F}_{2,t}$) of the cross-section of spreads.
2. Regress the second Principal Component on the specific local spread i under analysis.

$$\hat{F}_{2,t} = \gamma_i(Y_t^i - Y_t^{DE}) + \nu_{it}. \tag{A6.1}$$

3. Clean the second Principal Component from the information spanned by country i

$$\hat{F}_{2i,t} = \hat{F}_{2t} - [\hat{\gamma}_i(Y_t^i - Y_t^{DE})]. \tag{A6.2}$$

4. Obtain the external information as the average of foreign spreads.
5. Repeat 2 - 5 for each i.

$$\hat{F}_{2,it} = \sum_{i \neq j} w_{ij}(Y_t^j - Y_{t-1}^{DE}) + v_{it}. \tag{A6.3}$$

6. Estimate equation (A6.3) and retrieve the weights.

A6.2 Solution to GVAR Models

The *GVAR* framework employs country-specific ones built as weighted averages of foreign counterparts, where the weights capture the tightness of the interconnection between country i and country j as follows:

$$x_{it}^* = \sum_{j=0}^{N} w_{ij} x_{jt}, \quad w_{ii} = 0, \quad \sum_{j=0}^{N} w_{ij} = 1 \tag{A6.4}$$

resulting in the local *VARX** model.

$$\Phi_i(L, p_i)x_{it} = a_{i0} + a_{i1}t + \Upsilon_i(L, q_i)d_t + \Lambda_i(L, q_i)x_{it}^* + \nu_{it}, \tag{A6.5}$$

where p_i and q_i are the maximum lag considered for the local and foreign variables, respectively. The different coefficients are treated as unrestricted for estimation purposes.

Once local parameters in ((A6.5)) are estimated, we can simultaneously solve the model to obtain the *Global VAR* representation of the world. We now summarize the procedure as in Pesaran, Schuermann, and Weiner (2004).

Define the $(k_i + k_i^*) \times 1$ vector:

$$z_{i,t} = \begin{pmatrix} x_{i,t} \\ x_{i,t}^* \end{pmatrix} \tag{A6.6}$$

and rewrite ((A6.5)) as:

$$A_i z_{i,t} = B_i z_{i,t-1} + \epsilon_{it} \tag{A6.7}$$

with $A_i = (I_{k_i},\ -\Lambda_{i0})$ and $B_i = (\Phi_i,\ \Lambda_{i1})$. For ease of explanation, we here abstract, without loss of generality, from including the deterministic component, the time trend, and additional lags.

If we collect all the local variables in $x_t = (x_{0t}^{'}, x_{1t}^{'}, \ldots, x_{Nt}^{\prime})^{'}$ for $i = 1, \ldots, N$ with $k = \sum_1^N k_i$, we can rewrite:

$$z_{i,t} = W_i x_t. \tag{A6.8}$$

By substituting ((A6.8)) into ((A6.7)), we obtain:

$$A_i W_i x_t = B_i W_i x_{t-1} + \epsilon_{it}, \tag{A6.9}$$

Stacking all the equations together:

$$G x_t = H x_{t-1} + \epsilon_t, \tag{A6.10}$$

As in Candelon and Luisi (2020), by defining the interaction matrix $\tilde{W}$ as:

$$x_t^* = \tilde{W} x_t. \tag{A6.11}$$

We can express the matrices G and H as:

$$G = [I_k - \tilde{\Lambda}_0 \tilde{W}] \tag{A6.12}$$

$$H = [\tilde{\Phi} + \tilde{\Lambda}_1 \tilde{W}] \tag{A6.13}$$

Lastly, we can retrieve the $GVAR$ solution as:

$$x_t = G^{-1} H x_{t-1} + G^{-1} \epsilon_t. \tag{A6.14}$$

Therefore, the $GVAR$ solution is the reduced form VAR representation of the world. This representation is particularly useful for forecasting, scenario analysis, and the Generalized Impulse Response function.

A6.3 Bootstrap procedure for the GIRF confidence bounds

The advantages of the bootstrapping procedure when computing confidence intervals for impulse responses in the VAR framework are described in Kilian (1998). In our framework, to take into account also unknown forms of heteroskedasticity, we proposed the wild bootstrapped procedure as in Mammen (1993).

Specifically,

1. Estimate the local $VARX^*$ parameters in eq. ((6.2)) (after retrieving the weights as outlined in Appendix B), and the corresponding residuals u_t^i for each $i = 1, \ldots, N$.
2. Following the procedure outlined in Section A6.2, simultaneously solve the model to get the global VAR representation as of ((A6.14))

$$(Y_t - Y_t^{DE}) = G^{-1} H (Y_{t-1} - Y_{t-1}^{DE}) + G^{-1} u_t \tag{A6.15}$$

with

$$(Y_t - Y_t^{DE}) = ((Y_t^1 - Y_t^{DE})^{'}, (Y_t^2 - Y_t^{DE})^{'}, \ldots, (Y_t^N - Y_t^{DE})^{'})^{'}, \tag{A6.16}$$

$$G = [I_N - (diag(\beta_{12}, \beta_{22}, \ldots, \beta_{N2})\tilde{W})], \tag{A6.17}$$

$$H = [(diag(\beta_{11}, \beta_{21}, \ldots, \beta_{N1}) + (diag(\beta_{13}, \beta_{23}, \ldots, \beta_{N3})\tilde{W})], \tag{A6.18}$$

$$u_t = (u_t^{1^{\prime}}, u_t^{2^{\prime}}, \ldots, u_t^{N^{\prime}})^{'} \tag{A6.19}$$

3. Draw with replacement the sequence of errors $\{\tilde{u}_t\}_{t=2}^T = \{\tilde{u}_t^1, \ldots, \tilde{u}_t^N\}_{t=2}^T = \{k_t\hat{u}_t^1, \ldots, k_t\hat{u}_t^N\}_{t=2}^T$ with

$$k_t = \begin{cases} \frac{1+\sqrt{5}}{2}, & \text{with probability } p = \frac{\sqrt{5}-1}{2\sqrt{5}} \\ \frac{1-\sqrt{5}}{2}, & \text{with probability } 1-p. \end{cases}$$

   as in Mammen (1993), $\tilde{u}_t$ are the stacked estimated local errors.
4. Generate the bootstrapped data sample according to equation (A6.15). The initial values are the actual ones, and u_t is replaced by the bootstrapped $\tilde{u}_t$.
5. Estimate the local $VARX^*$ on the bootstrapped data sample and compute the corresponding Generalized Impulse Responses.
6. Repeat 3 - 5 a large number of times (in our example, we repeat the procedure $10,000$ times), and then extract the quantiles needed.

A6.4 Analysis of the interdependence factor

In this section, we investigate how the first *PC*, the interdependence factor, is associated with the observed variables that are often employed in the literature for the same purpose. Specifically, we consider the long term corporate *Baa-Aaa* spread to account for global risk aversion (Favero, 2013); the 10Y US treasury yield (Miranda-Agrippino and Rey, 2020); oil prices; equity market volatility; devaluation risk (De Santis, 2019) and the flight-to-liquidity factor (Monfort and Renne, 2013).

Table A6.1 shows the pairwise correlation between the first *PC* and the aforementioned observable global factors in the Pre-*ESD*, Post-*ESD* and COVID periods. While the negative sign of the correlations in the Post-*ESD* and COVID period reflects the negative loadings of the first *PC* in the respective samples, our unobserved global trend is strongly correlated with the average devaluation risk in the euro area. We can also notice that while oil prices are correlated with the global trend mainly in the Post-*ESD* period, flight-to-liquidity (KfW-Bund) and global risk aversion (Baa-Aaa) are strongly correlated with the first *PC* in the Pre-*ESD* and COVID period. Therefore, despite the intensity of correlations varies across the sub-samples we considered, the first *PC* is related to factors that capture devaluation risk, flight-to-liquidity mechanisms, and shifts global risk aversion, further supporting the use of the first *PC* as a global trend.

Table A6.1: First Principal Component: bilateral correlation with proxies of global factors.

Instrument	Pre-ESD	Post-ESD	COVID	Risk
VSTOXX	0.02	-0.08	-0.36	Equity Volatility
QUANTO CDS	0.63	-0.90	-0.71	Devaluation
KfW-Bund	0.79	0.13	-0.92	Flight-to-Liquidity
OIL-Brent	-0.31	-0.64	0.24	Inflation
OIL-WTI	-0.29	-0.70	0.32	Inflation
Baa-Aaa	0.81	-0.11	-0.67	Global Risk Aversion
US 10Y	-0.89	0.16	-0.19	Global Financial Cycle

Notes. This table reports the bilateral correlation between the first PC and the observable global factors in the Pre-ESD, Post-ESD, and COVID periods. The VSTOXX Indices are based on EURO STOXX 50 real-time options prices and are designed to reflect the market expectations of near-term up to long-term volatility by measuring the square root of the implied variance across all options of a given time to expiration. QUANTO CDS is defined as the average of German, Italian, French, and Spanish spreads between 5Y QUANTO CDS denominated in EUR and USD as in Favero and Missale (2016).
As in Monfort and Renne (2013), we consider the average spreads across maturities between bonds issued by a German agency (KfW) and their sovereign counterparts to identify liquidity-pricing effects. Indeed, KfW's liabilities are guaranteed by the German government, and therefore, these spreads are essentially liquidity-driven and can be used as a proxy of the flight-to-liquidity mechanism. OIL-Brent (DCOILBRENTEU) and Oil-WTI (DCOILWTICO) are Crude Oil Prices measured in US dollars per barrel.

Supplementary Material

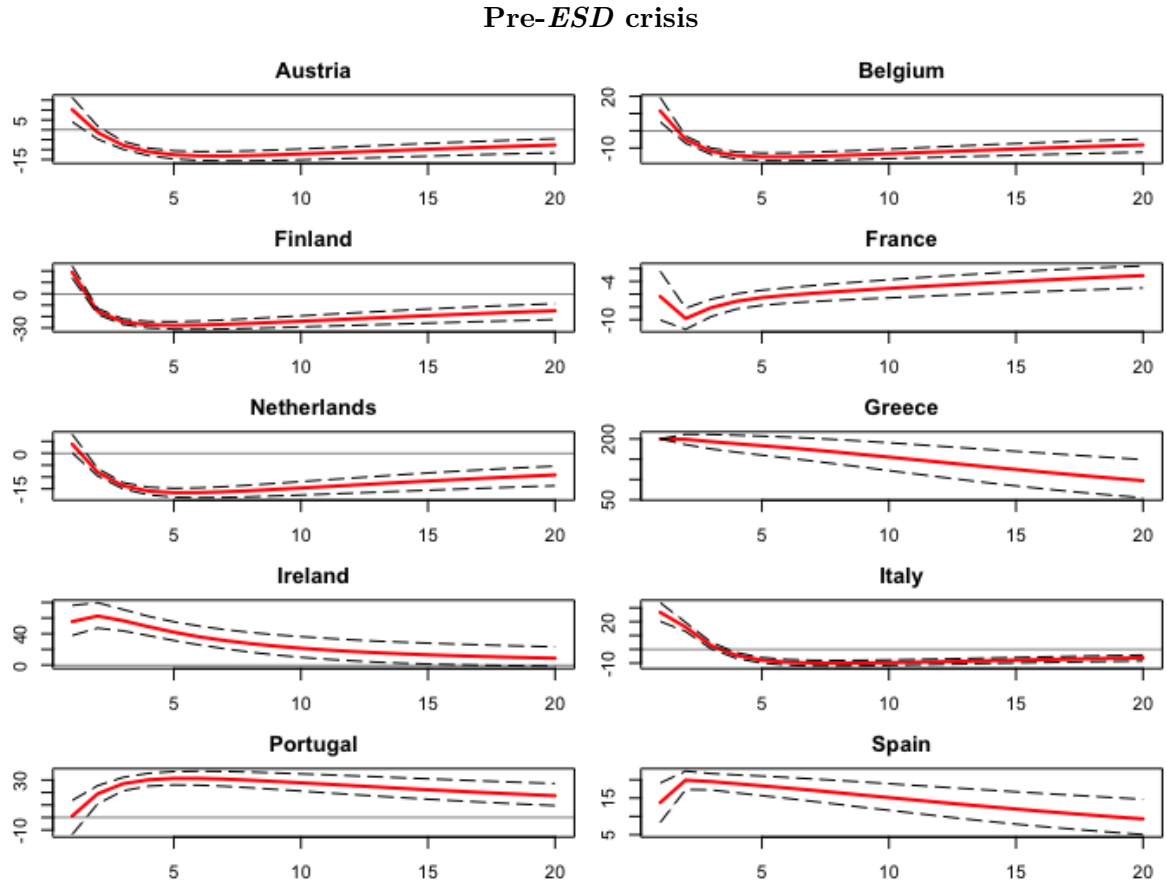


Figure S6.1: The figure reports the Generalized Impulse Response Functions to a 200 b.p. shock to the $10Y$ Greek Spread in the pre-crisis period. $[16^{th}, 84^{th}]$ confidence intervals are calculated using $10,000$ bootstrap replications as in A6.3.

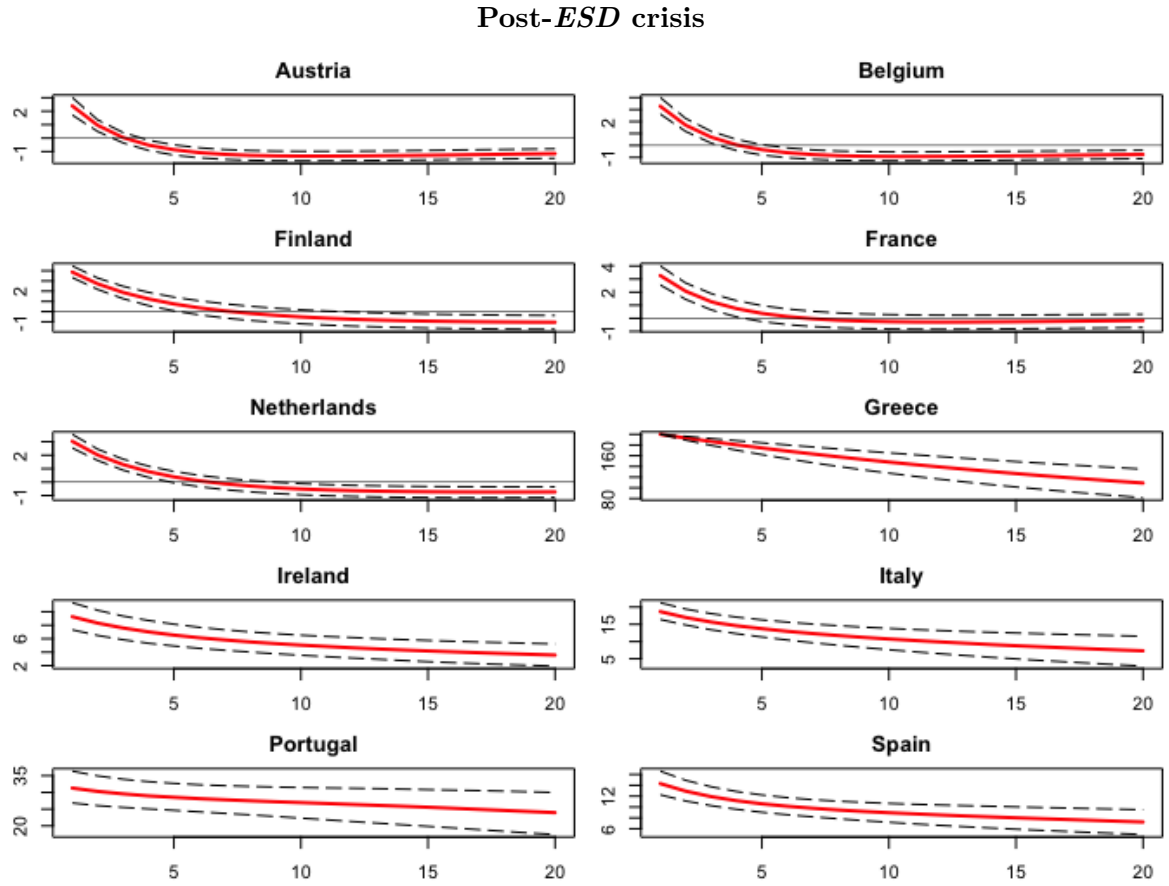


Figure S6.2: The figure reports the Generalized Impulse Response Functions to a 200 b.p. shock to the $10Y$ Greek Spread in the post-crisis period. $[16^{th}, 84^{th}]$ confidence intervals are calculated using $10,000$ bootstrap replications as in A6.3.

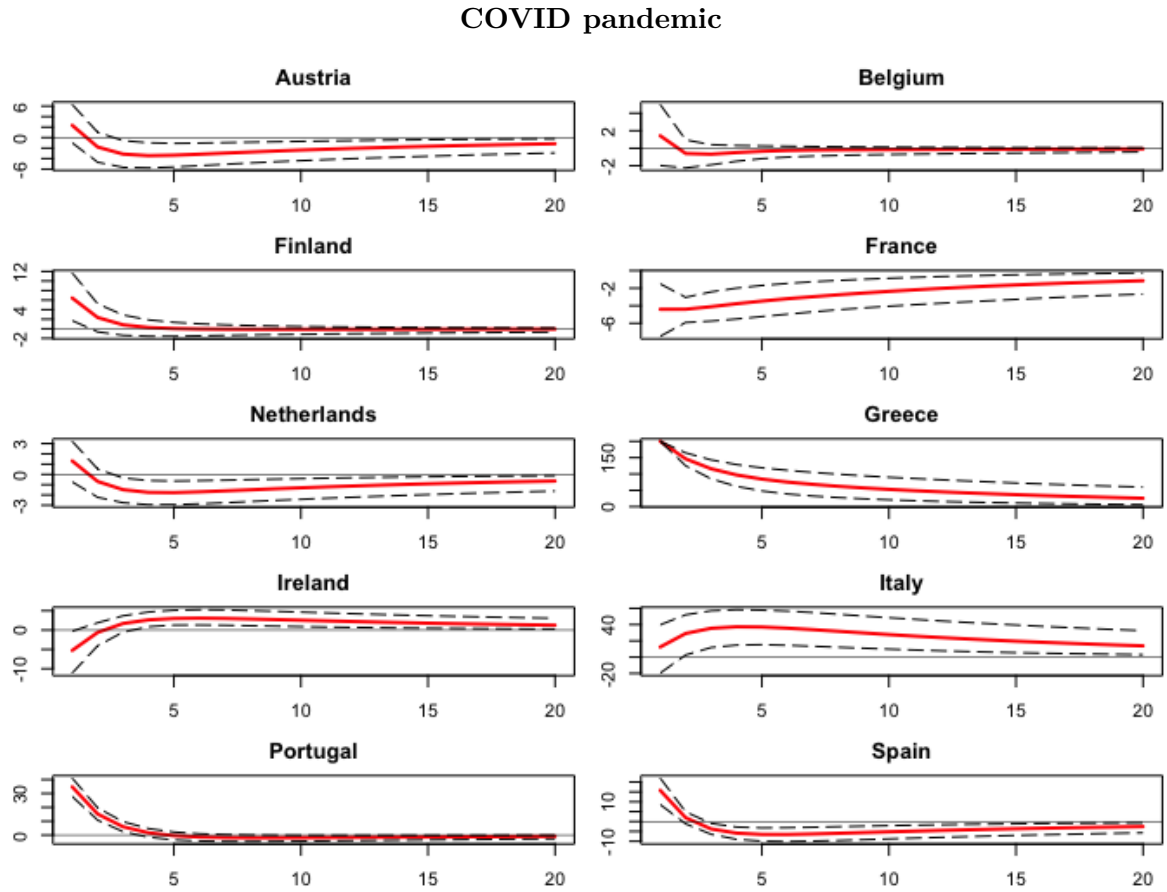


Figure S6.3: The figure reports the Generalized Impulse Response Functions to a 200 b.p. shock to the $10Y$ Greek Spread in the COVID pandemic period. $[16^{th}, 84^{th}]$ confidence intervals are calculated using $10,000$ bootstrap replications as in A6.3.

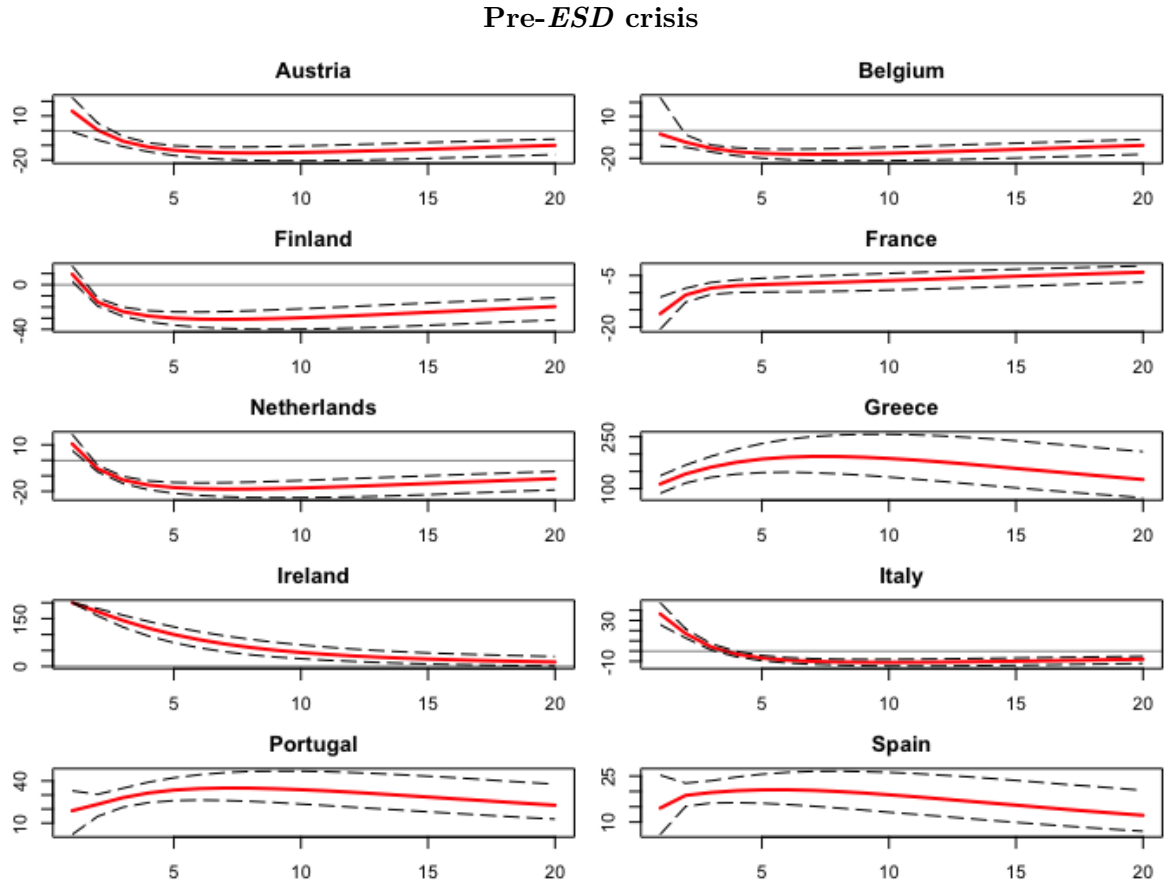


Figure S6.4: The figure reports the Generalized Impulse Response Functions to a 200 b.p. shock to the $10Y$ Irish Spread in the pre-crisis period. $[16^{th}, 84^{th}]$ confidence intervals are calculated using $10,000$ bootstrap replications as in A6.3.

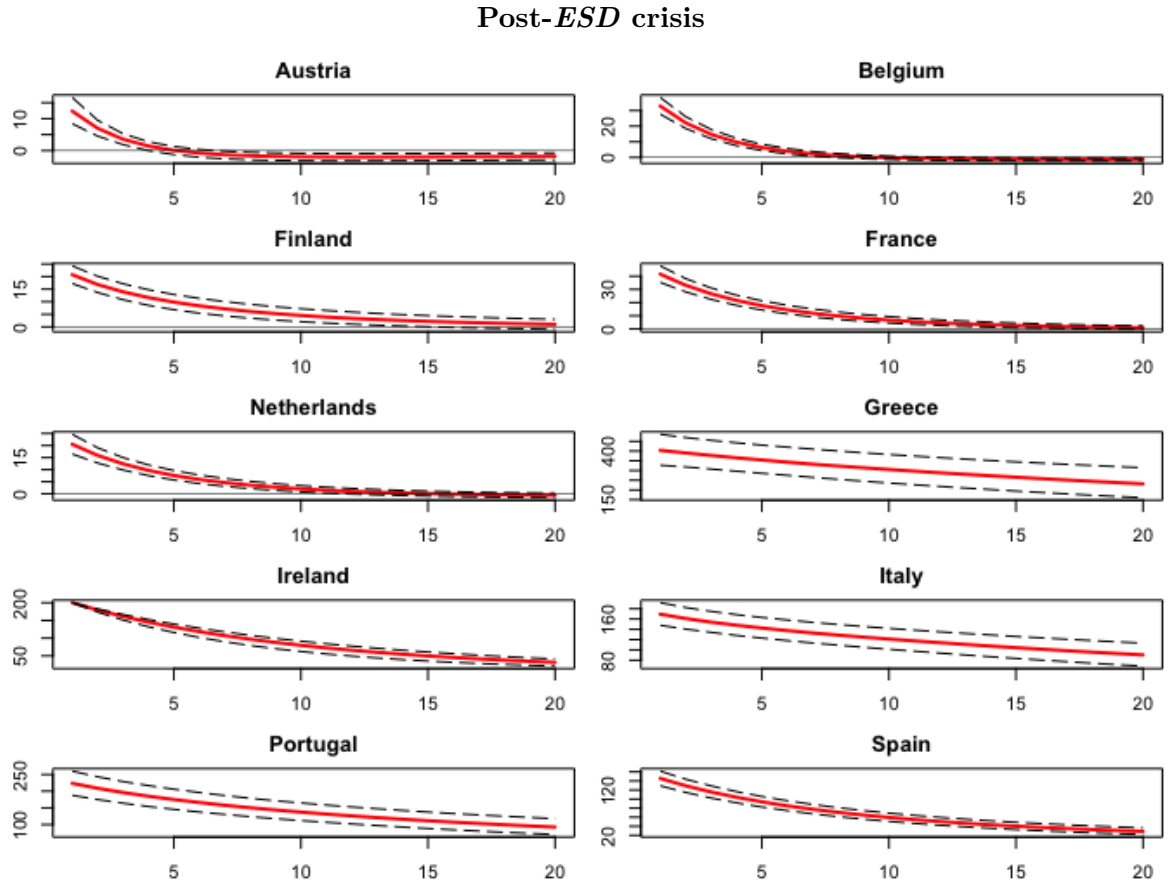


Figure S6.5: The figure reports the Generalized Impulse Response Functions to a 200 b.p. shock to the $10Y$ Irish Spread in the post-crisis period. $[16^{th}, 84^{th}]$ confidence intervals are calculated using $10,000$ bootstrap replications as in A6.3.

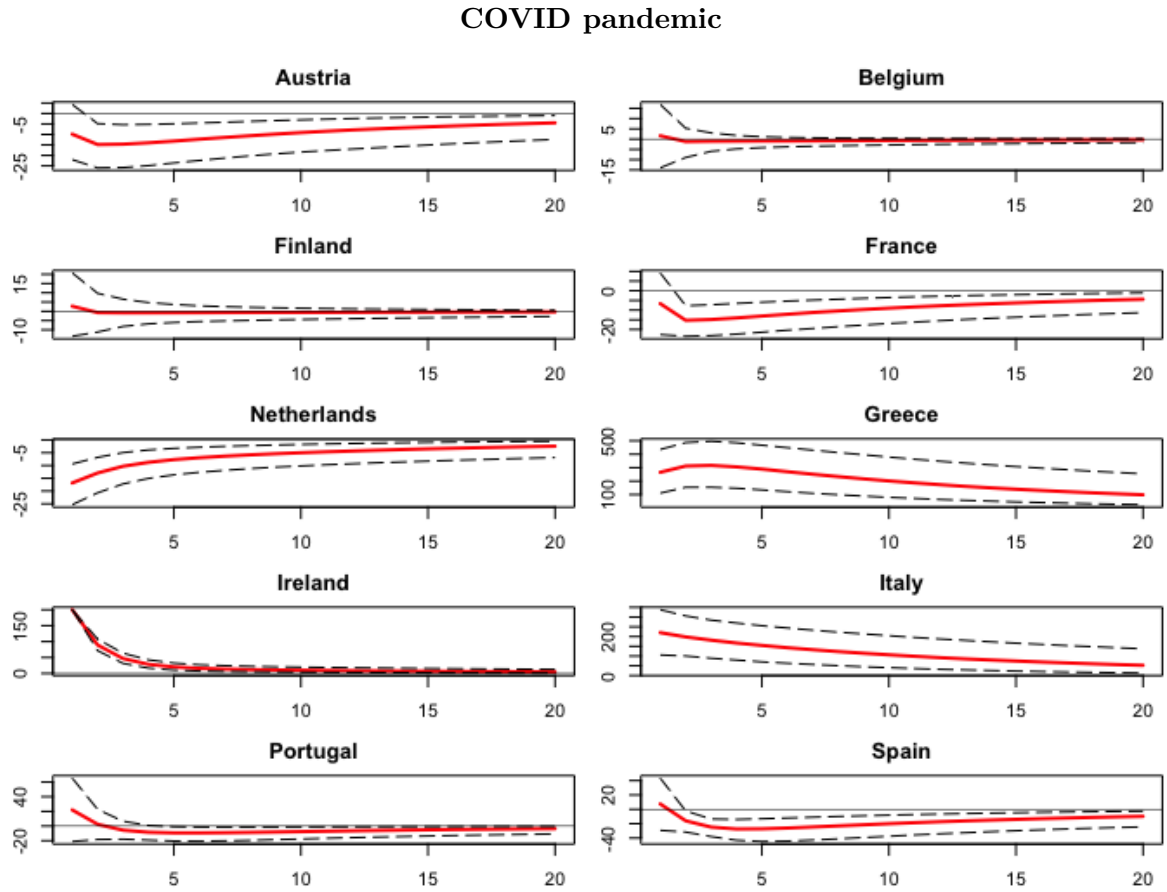


Figure S6.6: The figure reports the Generalized Impulse Response Functions to a 200 b.p. shock to the $10Y$ Irish Spread in the COVID pandemic period. $[16^{th}, 84^{th}]$ confidence intervals are calculated using $10,000$ bootstrap replications as in A6.3.

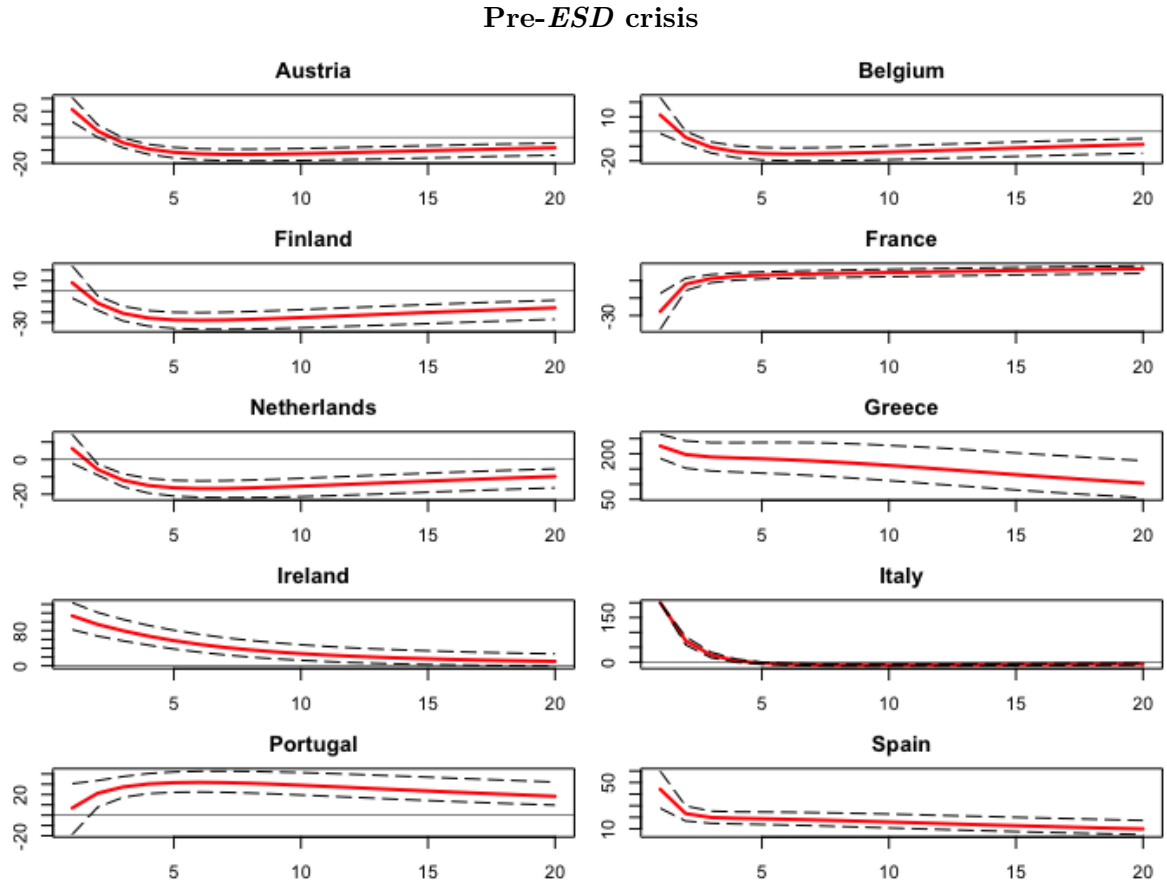


Figure S6.7: The figure reports the Generalized Impulse Response Functions to a 200 b.p. shock to the $10Y$ Italian Spread in the pre-crisis period. $[16^{th}, 84^{th}]$ confidence intervals are calculated using $10,000$ bootstrap replications as in A6.3.

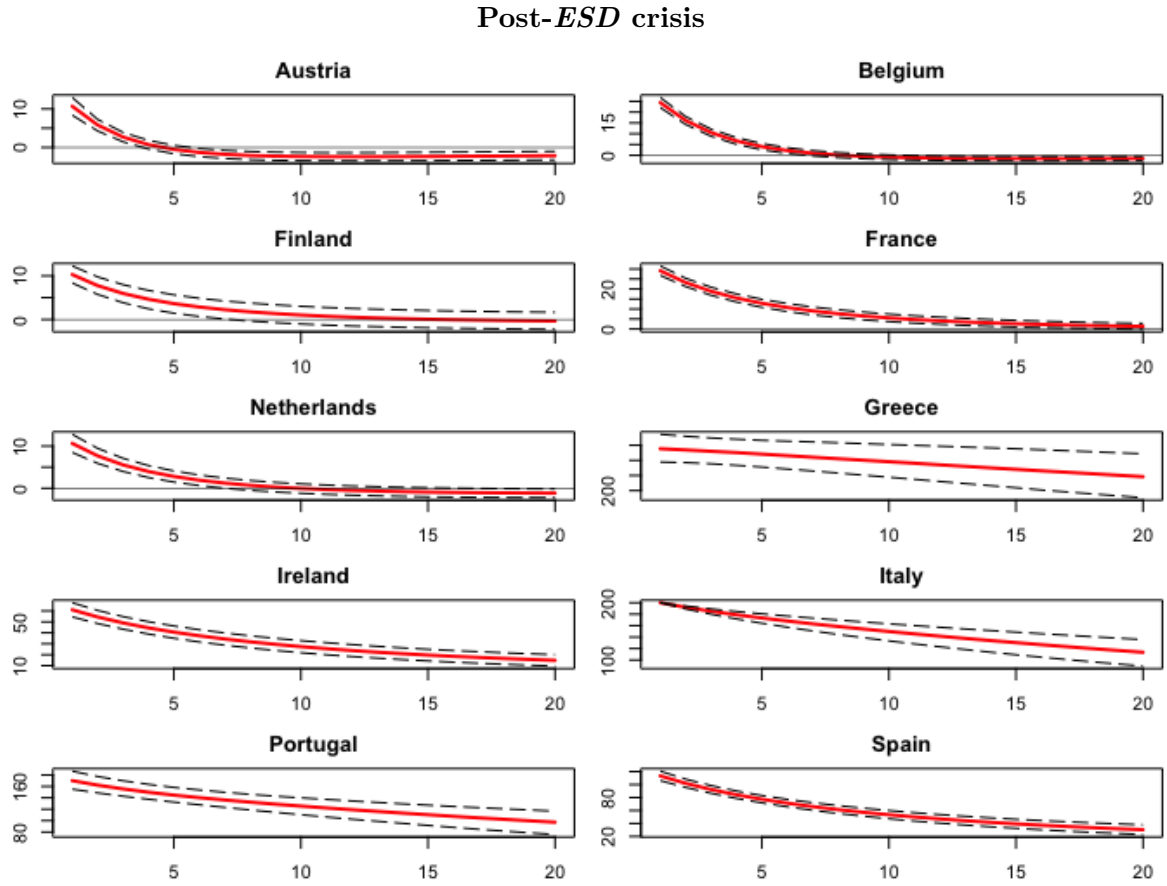


Figure S6.8: The figure reports the Generalized Impulse Response Functions to a 200 b.p. shock to the $10Y$ Italian Spread in the post-crisis period. [16^{th}, 84^{th}] confidence intervals are calculated using $10,000$ bootstrap replications as in A6.3.

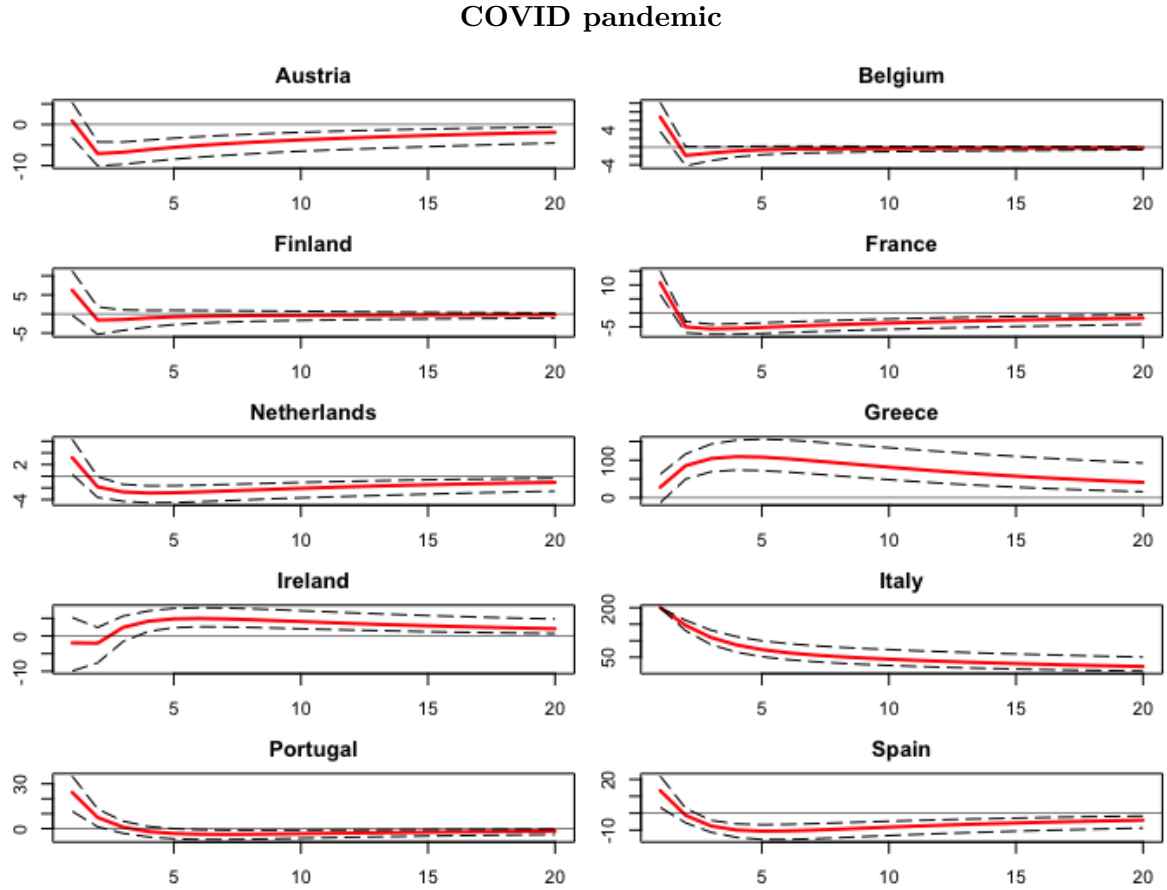


Figure S6.9: The figure reports the Generalized Impulse Response Functions to a 200 b.p. shock to the $10Y$ Italian Spread in the COVID pandemic period. $[16^{th}, 84^{th}]$ confidence intervals are calculated using $10,000$ bootstrap replications as in A6.3.

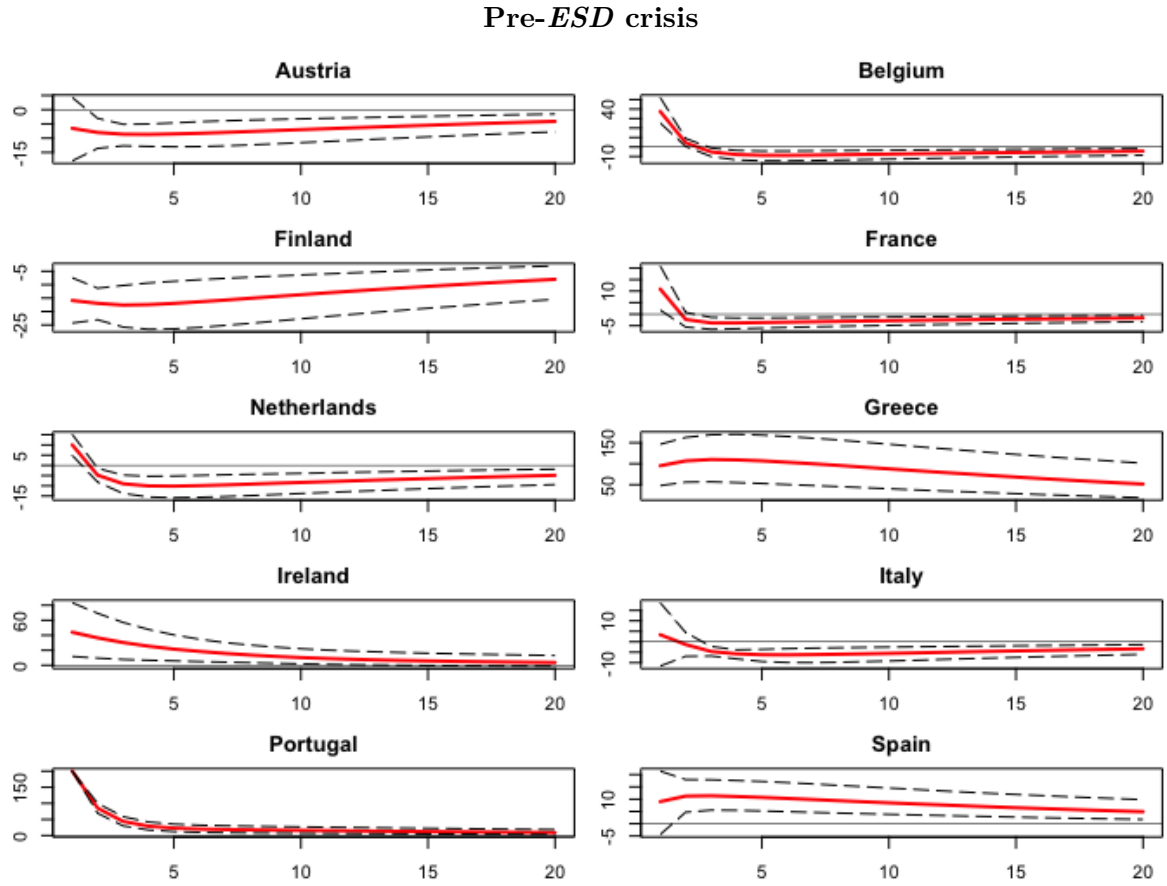


Figure S6.10: The figure reports the Generalized Impulse Response Functions to a 200 b.p. shock to the $10Y$ Portuguese Spread in the pre-crisis period. $[16^{th}, 84^{th}]$ confidence intervals are calculated using $10,000$ bootstrap replications as in A6.3.

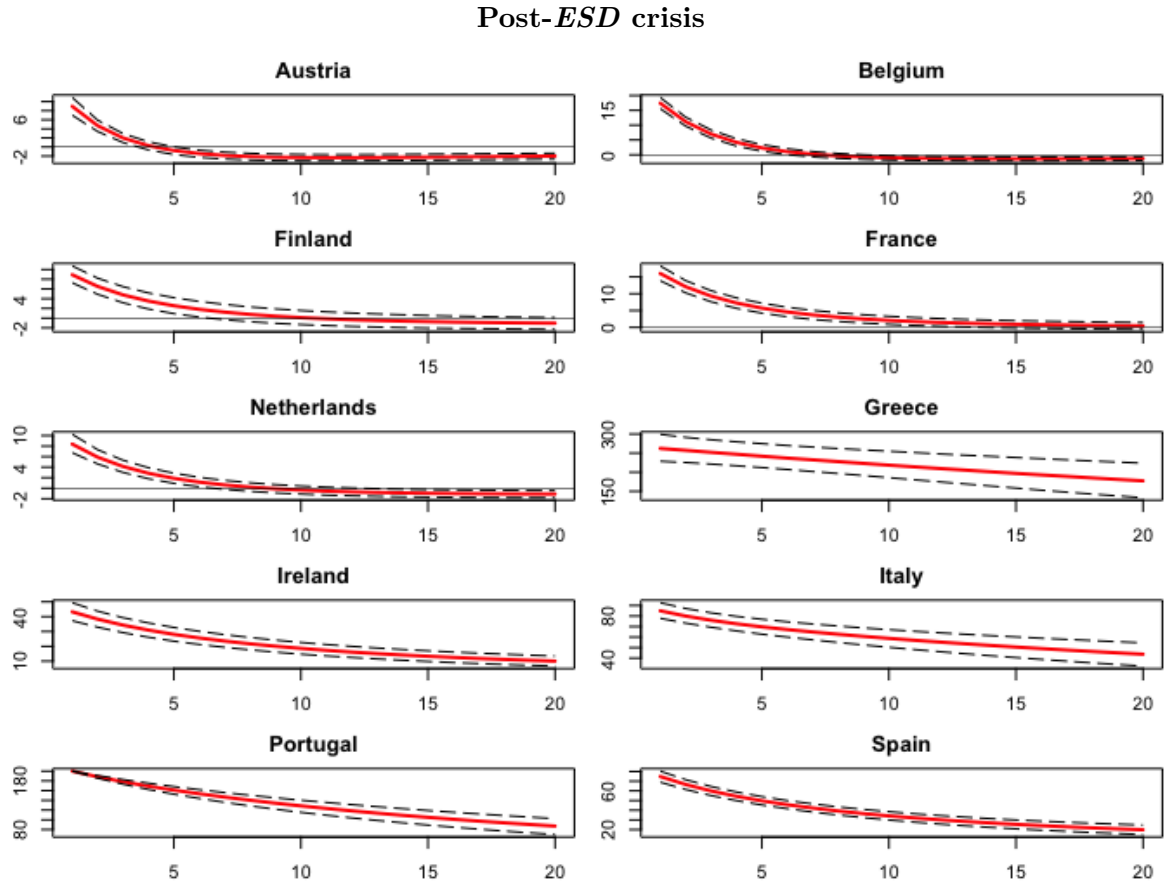


Figure S6.11: The figure reports the Generalized Impulse Response Functions to a 200 b.p. shock to the $10Y$ Portuguese Spread in the post-crisis period. [16^{th}, 84^{th}] confidence intervals are calculated using $10,000$ bootstrap replications as in A6.3.

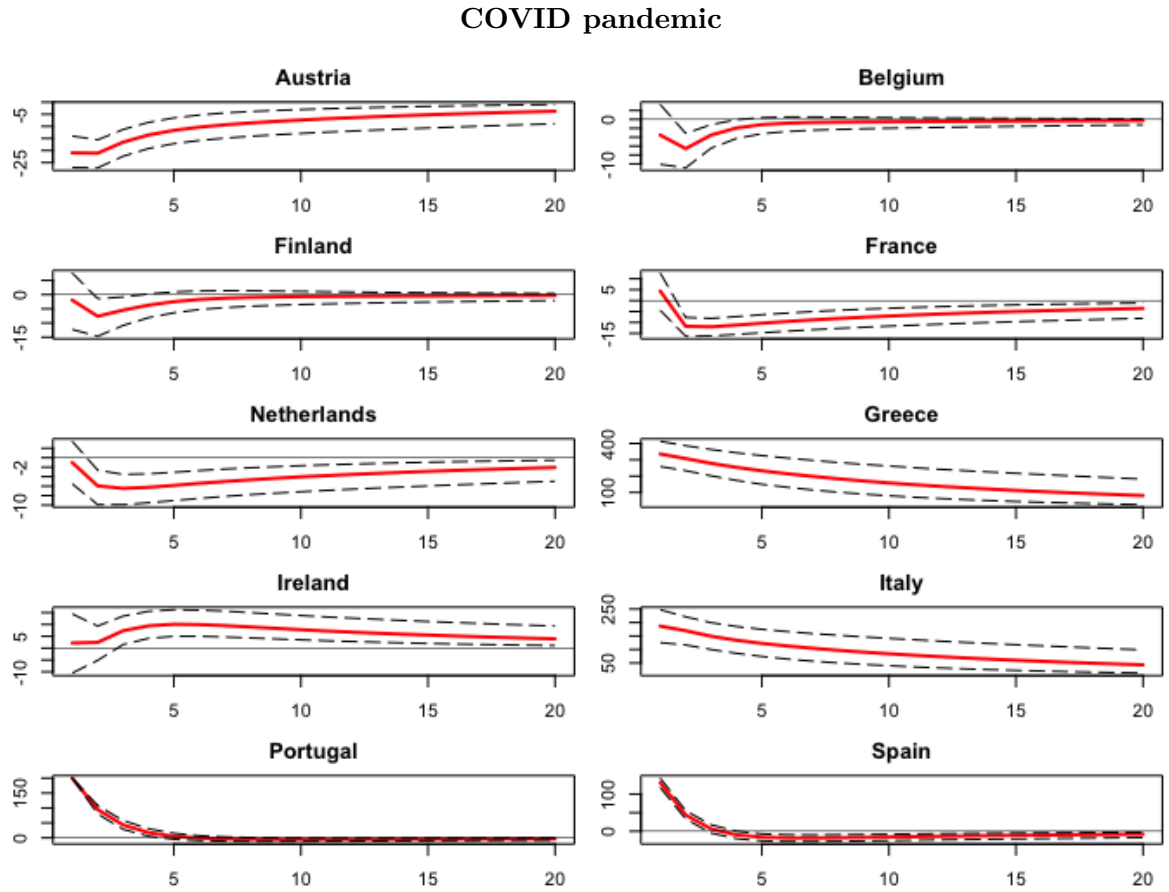


Figure S6.12: The figure reports the Generalized Impulse Response Functions to a 200 b.p. shock to the $10Y$ Portuguese Spread in the COVID pandemic period. [16^{th}, 84^{th}] confidence intervals are calculated using $10,000$ bootstrap replications as in A6.3.

Chapter 8

Conclusions and further research

Forecasting plays a crucial role in decision-making, particularly in economics and finance, where the optimal decisions depend on the future realization of the variable(s) under consideration. By utilizing forecasting techniques, it becomes possible to estimate future cash flows, earnings, and market trends, which can be used to evaluate investment opportunities, set financial goals, and make strategic decisions. With the advent of machine learning and big data, a wide range of predictions for the variable of interest has become readily available, resulting in a significant improvement in the accuracy of forecasting. This helps decision-makers make more informed decisions, especially when multiple forecasting methods suggest similar outcomes. However, individuals must ultimately make decisions, even when the various forecasting methods conflict. This can limit their ability to form and be confident in their expectations.

The aim of this thesis is to offer guidance to the forecasting community in dealing with uncertainty arising from the existence of multiple forecasts or situations where forecasts are not readily available or assessing predictive performance is not possible. In addition, this work sets the stage for further investigation in the field.

Chapter 2 presents techniques for combining forecasts through constrained optimization with a penalty. The chapter argues that combining forecasts offers advantages over forecast selection, and enforcing a non-negativity constraint on the weights and using penalties to measure the divergence from a reference combination scheme provides two benefits. First, it allows for forecast selection and combination to occur in a single step, potentially leading to sparse combining schemes. Second, it mitigates the risk of overfitting by shrinking an optimal but potentially unstable solution to a sub-optimal but satisfactory one. These techniques offer a way to effectively combine forecasts and improve forecasting accuracy. Chapter 5 connects optimal combining weights to the field of forecast evaluation, introducing a novel forecast encompassing test. By analyzing euro area inflation projections from various sources, the chapter shows that the European Central Bank (ECB) is the most informative forecaster overall, although its weight varies over time due to changes in macro-financial conditions and monetary policy actions. The findings suggest that optimal combining weights can be a useful tool for evaluating monetary policy.

In the credit industry, practitioners and regulators are also often confronted with the uncertainty deriving from model risk. This is the risk associated with the potential discrepancy between the expected and unexpected results obtained when confronting a model and with the actual outcomes. Usual questions affecting model risk are what sampling strategy, what statistical framework, and what predictors we should consider when designing a model to assess credit risk. In Chapter 3, we reveal that the most

promising approach to forecast corporate bonds recovery rate involves combining models that are trained on security-specific characteristics and a limited number of well-established and theoretically sound recovery rate determinants, including measures of financial and macroeconomic uncertainty. In Chapter 4, we focus on the predictor set and we challenge the previous research suggesting that only contact-specific variables such as exposure-at-default or age were relevant, while macro variables like the unemployment rate or business cycle were irrelevant. Instead, by analyzing a larger dataset of over 6 million Italian consumer debts from 2007 to 2019, considering more than 40 predictors, and using a variety of models including standard regressions, machine learning techniques, and forecast combination schemes, we find that regional social and welfare indicators and credit cycle variables are important predictors and can enhance forecasting performance by up to 6 p.p. in terms of R^2.

In the previous chapters, we have assumed that the forecaster is confronted with multiple forecasts. In Chapter 6, we instead consider the case in which the decision maker is confronted with the opposite scenarios: no forecasts are readily available and/or tracking historical forecast errors to assess predictive performance is not viable. Specifically, Chapter 6 examines how an unexpected rise in the yield of a public debt issued by one of the peripheral countries in the Eurozone would impact the yields of other participating countries.

The objective of this work was to equip regulators, policymakers, and individuals with tools that can help rebuild confidence in the process of forming expectations and reduce the impact of uncertainty on decision-making. Clearly, there is still much to do and, in terms of proposals for further research, we foresee three main directions.

First, throughout this work, we have implicitly assumed that forecasters aim to minimize the variance of the forecast error by focusing on mean point forecasts, and optimal and robust combining techniques have been derived with respect to square loss functions. However, decision-makers may have different preferences and priorities, such as preferring conservative earning forecasts rather than being overconfident during uncertain times. Moreover, decision-makers should also consider rare and extreme events when evaluating risk and developing risk management strategies. The questions of how to tailor forecast combination and evaluation techniques to the preference of the individual forecaster and how to handle extreme events remain open. Portfolio optimization techniques with a focus on higher moments and conditional quantile functions can offer new insights into how to optimally combine and evaluate predictive strategies when departing from the mean square loss.

Second, when dealing with economic and financial time series, the likelihood of being confronted with structural breaks is high. In the thesis, we handled this issue by estimating the quantities of interest employing rolling windows. However, combining weights remain inherently fixed within each pocket. Explicitly modeling time-varying weights exploiting cross-sectional information may offer new insights on when a forecasting method offers superior performance. One solution may come from *location-scale* regressions that jointly model the mean and scale (variance) of the forecasts, and consequently forecast errors.

Finally, this thesis also focused on predicting the potential spillover effects of a shock across interconnected markets. Global VARs and Global VECMs are the *go-to* frameworks to model interconnections. In asset pricing, they might offer the advantage of escaping *factor zoo* problem from the root as they model assets as a function of assets and not specific factors. However, they still rely on the choice of a weighting matrix $\boldsymbol{W}$ that measures foreign exposures. However, this only measures cross-sectional dependence, and no information on the *direction* of the spillover can be extrapolated. Studies building on directed graphical models may fill this gap, helping to better understand the direction of the spillovers.

We leave these research directions for future work.

BIBLIOGRAPHY

Acharya, V. V., S. T. Bharath, and A. Srinivasan (2007): "Does industry-wide distress affect defaulted firms? Evidence from creditor recoveries," *Journal of Financial Economics*, 85(3), 787–821.

Aiolfi, M., and C. A. Favero (2005): "Model uncertainty, thick modelling and the predictability of stock returns," *Journal of Forecasting*, 24(4), 233–254.

Al-Eyd, A. J., and P. Berkmen (2013): "Fragmentation and Monetary Policy in the Euro Area," *IMF Working Papers*, 13(208), 1.

Alesina, A., and R. Gatti (1995): "Independent Central Banks: Low Inflation at No Cost?," *The American Economic Review*, 85(2), 196–200.

Alexopoulos, M., and J. Cohen (2015): "The power of print: uncertainty shocks, markets, and the economy," *International Review of Economics & Finance*, 40, 8–28.

Altman, E., B. Brady, A. Resti, and A. Sironi (2005): "The link between default and recovery rates: theory, empirical evidence, and implications," *The Journal of Business*, 78(6), 2203–2227.

Altman, E., and B. Karlin (2009): "The re-emergence of distressed exchanges in corporate restructurings," *Journal of Credit Risk*, 5, 43–56.

Altman, E., and V. Kishore (1996): "Almost everything you wanted to know about recoveries on defaulted bonds," *Financial Analysts Journal*, Nov.-Dec., 57–64.

Altman, E. I., and E. A. Kalotay (2014): "Ultimate recovery mixtures," *Journal of Banking & Finance*, 40, 116–129.

Amendola, A., M. Braione, V. Candila, and G. Storti (2020): "A Model Confidence Set approach to the combination of multivariate volatility forecasts," *International Journal of Forecasting*, 36(3), 873–891.

Andersen, L., and J. Sidenius (2004): "Extensions to the Gaussian copula: random recovery and random factor loadings," *Journal of Credit Risk*, 1, 29–70.

Andrade, P., and H. Le Bihan (2013): "Inattentive professional forecasters," *Journal of Monetary Economics*, 60(8), 967–982.

Ang, A., G. Bekaert, and M. Wei (2007): "Do macro variables, asset markets, or surveys forecast inflation better?," *Journal of Monetary Economics*, 54(4), 1163–1212.

Apley, D. W., and J. Zhu (2020): "Visualizing the effects of predictor variables in black box supervised learning models," *Journal of the Royal Statistical Society Series B*, 82(4), 1059–1086.

Ardia, D., K. Boudt, and J.-P. Gagnon-Fleury (2017): "RiskPortfolios: Computation of risk-based portfolios in R," *Journal of Open Source Software*, 10(2).

Aruoba, S. B., F. X. Diebold, J. Nalewaik, F. Schorfheide, and D. Song (2012): "Improving U.S. GDP Measurement: A Forecast Combination Perspective," in *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis*, pp. 1–25. Springer New York.

Atiya, A. F. (2020): "Why does forecast combination work so well?," *International Journal of Forecasting*, 36(1), 197–200.

Bachmann, R., S. Elstner, and E. Sims (2013): "Uncertainty and economic activity: evidence from business survey data," *American Economic Journal: Macroeconomics*, 5(2), 217–249.

Baele, L., A. Ferrando, P. Hördahl, E. Krylova, and C. Monnet (2004): "Measuring European financial integration," *Oxford Review of Economic Policy*, 20(4), 509–530.

Bailey, N., S. Holly, and M. H. Pesaran (2016): "A two-stage approach to spatio-temporal analysis with strong and weak cross-sectional dependence," *Journal of Applied Econometrics*, 31(1), 249–280.

Bailey, N., G. Kapetanios, and M. H. Pesaran (2016): "Exponent of cross-sectional dependence: Estimation and inference," *Journal of Applied Econometrics*, 31(6), 929–960.

Baker, S. R., N. Bloom, and S. J. Davis (2016): "Measuring Economic Policy Uncertainty," *The Quarterly Journal of Economics*, 131(4), 1593–1636.

Balabanova, Z., and R. Brüggemann (2017): "External information and monetary policy transmission in new EU member states: results from FAVAR models," *Macroeconomic Dynamics*, 21(02), 311–335.

Ball, L. (1992): "Why does high inflation raise inflation uncertainty?," *Journal of Monetary Economics*, 29(3), 371–388.

Bank of Italy (2020): "L'economia delle regioni italiane. Dinamiche recenti e aspetti strutturali," Discussion paper.

Bastos, J. (2014): "Ensemble Predictions of Recovery Rates," *Journal of Financial Services Research*, 46, 177–193.

BCBS (2006): "Basel II: international convergence of capital measurement and capital standards: a revised framework - comprehensive version," .

——— (2011): "Basel III: a global regulatory framework for more resilient banks and banking systems," .

——— (2017): "Basel III: finalising post-crisis reforms," Available at: `https://www.bis.org/bcbs/publ/d424.htm`.

Beber, A., M. W. Brandt, and K. A. Kavajecz (2009): "Flight-to-quality or flight-to-liquidity? Evidence from the euro-area bond market," *The Review of Financial Studies*, 22(3), 925–957.

Beck, T., J. Grunert, W. Neus, and A. Walter (2017): "What Determines Collection Rates of Debt Collection Agencies?," *Financial Review*, 52(2), 259–279.

Behr, P., A. Guettler, and F. Miebs (2013): "On portfolio optimization: Imposing the right constraints," *Journal of Banking & Finance*, 37(4), 1232–1242.

Bekaert, G., M. Hoerova, and M. Lo Duca (2013): "Risk, uncertainty and monetary policy," *Journal of Monetary Economics*, 60(7), 771–788.

Bellotti, A., D. Brigo, P. Gambetti, and F. Vrins (2021): "Forecasting recovery rates on non-performing loans with machine learning," *International Journal of Forecasting*, 37(1), 428–444.

Bellotti, T., and J. Crook (2012): "Loss given default models incorporating macroeconomic variables for credit cards," *International Journal of Forecasting*, 28(1), 171–182.

Ben-Porat, O., and M. Tennenholtz (2017): "Best Response Regression," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, p. 1498–1507, Red Hook, NY, USA. Curran Associates Inc.

Berd, A. (2005): "Recovery swaps," *Journal of Credit Risk*, 1(3), 61–70.

Bernanke, B. S. (1986): "Alternative explanations of the money-income correlation," *Carnegie-Rochester Conference Series on Public Policy*, 25, 49–99.

Bernanke, B. S., J. Boivin, and P. Eliasz (2005): "Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach," *The Quarterly journal of economics*, 120(1), 387–422.

Bernanke, B. S., and F. S. Mishkin (1997): "Inflation Targeting: A New Framework for Monetary Policy?," *Journal of Economic Perspectives*, 11(2), 97–116.

Betz, J., R. Kellner, and D. Rösch (2018): "Systematic Effects among Loss Given Defaults and their Implications on Downturn Estimation," *European Journal of Operational Research*, 271(3), 1113–1144.

Betz, J., S. Krüger, R. Kellner, and D. Rösch (2020): "Macroeconomic effects and frailties in the resolution of non-performing loans," *Journal of Banking & Finance*, 112, 105212.

Blanchard, O. J., and D. Quah (1989): "The Dynamic Effects of Aggregate Demand and Supply Disturbances," *American Economic Review*, 79(4), 655–673.

Bloom, N. (2009): "The impact of uncertainty shocks," *Econometrica*, 77(3), 623–685.

Bowley, A. L. (1920): *Elements of Statistics*. New York: Charles Scribner's Sons.

Breiman, L. (1996a): "Bagging Predictors," *Machine Learning*, 24(2), 123–140.

——— (1996b): "Stacked regressions," *Machine Learning*, 24(1), 49–64.

——— (2001): "Random forests," *Machine learning*, 45(1), 5–32.

Breiman, L., J. Friedman, R. Olshen, and C. Stone (1984): *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA.

Breiman, L., and J. H. Friedman (1985): "Estimating Optimal Transformations for Multiple Regression and Correlation," *Journal of the American Statistical Association*, 80(391), 580–598.

Bris, A., I. Welch, and N. Zhu (2006): "The costs of bankruptcy: chapter 7 liquidation versus chapter 11 reorganization," *The Journal of Finance*, 61, 1253–1303.

Brodie, J., I. Daubechies, C. De Mol, D. Giannone, and I. Loris (2009): "Sparse and stable Markowitz portfolios," *Proceedings of the National Academy of Sciences*, 106(30), 12267–12272.

Bruche, M., and C. González-Aguado (2010): "Recovery rates, default probabilities and the credit cycle," *Journal of Banking & Finance*, 34(4), 754–764.

Bunn, D. W. (1981): "Two Methodologies for the Linear Combination of Forecasts," *Journal of the Operational Research Society*, 32(3), 213–222.

——— (1985): "Statistical efficiency in the linear combination of forecasts," *International Journal of Forecasting*, 1(2), 151–163.

Busetti, F., and J. Marcucci (2013): "Comparing forecast accuracy: A Monte Carlo investigation," *International Journal of Forecasting*, 29(4), 13–27.

Candelon, B., and A. Luisi (2020): "Testing for the Validity of W in GVAR models," Discussion paper, Université catholique de Louvain, Louvain Finance (LFIN).

Candelon, B., A. Luisi, and F. Roccazzella (2022): "Fragmentation in the European Monetary Union: Is it really over?," *Journal of International Money and Finance*, 122, 102545.

Candelon, B., and S. Tokpavi (2016): "A non-parametric test for Granger-causality in distribution with application to financial contagion," *Journal of Business & Economic Statistics*, 34(2), 240–253.

Capistrán, C. (2008): "Bias in Federal Reserve inflation forecasts: Is the Federal Reserve irrational or just cautious?," *Journal of Monetary Economics*, 55(8), 1415–1427.

Caputo, B., K. L. Sim, F. Furesjo, and A. J. Smola (2002): "Appearance-based Object Recognition using SVMs: Which Kernel Should I Use?," in *Proc. of NIPS workshop on Statistical methods for computational experiments in visual processing and computer vision*.

Caruana, R., and A. Niculescu-Mizil (2006): "An Empirical Comparison of Supervised Learning Algorithms," in *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, p. 161–168, New York, NY, USA. Association for Computing Machinery.

Caruana, R., A. Niculescu-Mizil, G. Crew, and A. Ksikes (2004): "Ensemble Selection from Libraries of Models," in *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04, p. 18, New York, NY, USA. Association for Computing Machinery.

Chen, T., and C. Guestrin (2016): "XGBoost," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM.

Cheung, Y.-W., S. Steinkamp, and F. Westermann (2020): "Capital flight to Germany: Two alternative measures," *Journal of International Money and Finance*, 102, 102095.

Claeskens, G., J. R. Magnus, A. L. Vasnev, and W. Wang (2016): "The forecast combination puzzle: A simple theoretical explanation," *International Journal of Forecasting*, 32(3), 754–762.

Clark, T. E., and M. W. McCracken (2001): "Tests of equal forecast accuracy and encompassing for nested models," *Journal of Econometrics*, 105(1), 85–110, Forecasting and empirical methods in finance and macroeconomics.

Clements, M., and D. Hendry (1998): *Forecasting Economic Time Series*. Cambridge University Press.

Cole, H. L., and T. J. Kehoe (2000): "Self-Fulfilling Debt Crises," *The Review of Economic Studies*, 67(1), 91–116.

Conflitti, C., C. D. Mol, and D. Giannone (2015): "Optimal combination of survey forecasts," *International Journal of Forecasting*, 31(4), 1096–1103.

Copelovitch, M., J. Frieden, and S. Walter (2016): "The political economy of the Euro crisis," *Comparative Political Studies*, 49(7), 811–840.

Corsetti, G., and L. Dedola (2016): "The Mystery of the Printing Press: Monetary Policy and Self-Fulfilling Debt Crises," *Journal of the European Economic Association*, 14(6), 1329–1371.

Corsetti, G., J. B. Duarte, and S. Mann (2020): "The heterogeneous transmission of ECB policies: Lessons for a post-COVID Europe," *VoxEu*.

Corsetti, G., K. Kuester, A. Meier, and G. J. Muller (2014): "Sovereign risk and belief-driven fluctuations in the euro area," *Journal of Monetary Economics*, 61, 53–73, Carnegie-Rochester-NYU Conference Series on "Fiscal Policy in the Presence of Debt Crises" held at the Stern School of Business, New York University on April 19-20, 2013.

Davydenko, S. A., and J. R. Franks (2008): "Do Bankruptcy Codes Matter? A Study of Defaults in France, Germany, and the U.K," *Journal of Finance*, 63(2), 565–608.

De Grauwe, P., and Y. Ji (2012): "Mispricing of Sovereign Risk and Macroeconomic Stability in the Eurozone," *Journal of Common Market Studies*, 50(6), 866–880.

De Grauwe, P., and Y. Ji (2022): "The fragility of the Eurozone: Has it disappeared?," *Journal of International Money and Finance*, 120, 102546.

De Santis, R. (2019): "Redenomination Risk," *Journal of Money, Credit and Banking*, 51(8), 2173–2206.

DeMiguel, V., L. Garlappi, F. J. Nogales, and R. Uppal (2009): "A Generalized Approach to Portfolio Optimization: Improving Performance by Constraining Portfolio Norms," *Management Science*, 55(5), 798–812.

DeMiguel, V., L. Garlappi, and R. Uppal (2007): "Optimal Versus Naive Diversification: How Inefficient is the 1/NPortfolio Strategy?," *Review of Financial Studies*, 22(5), 1915–1953.

Dickinson, J. P. (1973): "Some Statistical Results in the Combination of Forecasts," *Journal of the Operational Research Society*, 24(2), 253–260.

Diebold, F. X. (1989): "Forecast combination and encompassing: Reconciling two divergent literatures," *International Journal of Forecasting*, 5(4), 589–592.

Diebold, F. X., and R. S. Mariano (1995): "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics*, 13(3), 253–263.

Diebold, F. X., and P. Pauly (1990): "The use of prior information in forecast combination," *International Journal of Forecasting*, 6(4), 503–508.

Diebold, F. X., and M. Shin (2019): "Machine learning for regularized survey forecast combination: Partially-egalitarian LASSO and its derivatives," *International Journal of Forecasting*, 35(4), 1679–1691.

Dodge, S., and L. Karam (2016): "Understanding how image quality affects deep neural networks," in *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6.

Dong, X., L. Yan, D. Rapach, and Z. Guofu (2020): "Anomalies and the expected market return," Working Paper.

Dovern, J., U. Fritsche, and J. Slacalek (2012): "Disagreement Among Forecasters in G7 Countries," *The Review of Economics and Statistics*, 94(4), 1081–1096.

Easaw, J., and R. Golinelli (2021): "Professionals Inflation Forecasts: The Two Dimensions Of Forecaster Inattentiveness," *Oxford Economic Papers*, gpab012.

ECB (2009): "Uncertainty and the economic prospects for the euro area," *ECB Economic Bulletin*, pp. 58–61.

——— (2016): "The impact of uncertainty on activity in the euro area," *ECB Economic Bulletin*, pp. 55–74.

Ehrbeck, T., and R. Waldmann (1996): "Why are Professional Forecasters Biased? Agency Versus Behavioral Explanations," *The Quarterly Journal of Economics*, 111(1), 21–40.

Elhorst, J. P., M. Gross, and E. Tereanu (2020): "Cross-sectional dependence and spillovers in space and time: where spatial econometrics and global var models meet," *Journal of Economic Surveys*.

Estrella, A., and F. S. Mishkin (1997): "The predictive power of the term structure of interest rates in Europe and the United States: Implications for the European Central Bank," *European Economic Review*, 41(7), 1375–1401.

Fan, J., Y. Liao, and M. Mincheva (2013): "Large covariance estimation by thresholding principal orthogonal complements," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 75(4), 603–680.

Fan, J., J. Zhang, and K. Yu (2012): "Vast Portfolio Selection With Gross-Exposure Constraints," *Journal of the American Statistical Association*, 107(498), 592–606.

Faust, J., and J. H. Wright (2013): "Chapter 1 - Forecasting Inflation," in *Handbook of Economic Forecasting*, ed. by G. Elliott, and A. Timmermann, vol. 2 of *Handbook of Economic Forecasting*, pp. 2–56. Elsevier.

Favero, C., and A. Missale (2012): "Sovereign spreads in the eurozone: which prospects for a Eurobond?," *Economic Policy*, 27(70), 231–273.

Favero, C. A. (2013): "Modelling and forecasting government bond spreads in the euro area: a GVAR model," *Journal of Econometrics*, 177(2), 343–356.

Favero, C. A., and A. Missale (2016): "Contagion in the EMU–the role of Eurobonds with OMTs," *Review of Law & Economics*, 12(3), 555–584.

Fedaseyeu, V. (2020): "Debt collection agencies and the supply of consumer credit," *Journal of Financial Economics*, 138(1), 193–221.

Ferrari, S., and F. Cribari-Neto (2004): "Beta Regression for Modelling Rates and Proportions," *Journal of Applied Statistics*, 31(7), 799–815.

Forbes, K. J., and R. Rigobon (2002): "No contagion, only interdependence: measuring stock market comovements," *The Journal of Finance*, 57(5), 2223–2261.

François, P. (2019): "The Determinants of market-implied recovery rates," *Risks*, 7(2), 1–15.

Franks, J. R., and W. N. Torous (1994): "A comparison of financial recontracting in distressed exchanges and Chapter 11 reorganizations," *Journal of Financial Economics*, 35(3), 349 – 370.

Friedman, J., T. Hastie, R. Tibshirani, N. Balasubramanian, and S. Noah (2019): *glmnet: Lasso and elastic-net regularized generalized linear models*R package version 3.0-2.

Friedman, J. H. (1991): "Multivariate Adaptive Regression Splines," *The Annals of Statistics*, 19(1), 1–67.

——— (2002): "Stochastic gradient boosting," *Computational Statistics & Analysis*, 38(4), 367–378.

Fu, W., and K. Knight (2000): "Asymptotics for lasso-type estimators," *The Annals of Statistics*, 28(5), 1356 – 1378.

Gabrieli, S., and C. Labonne (2018): "Bad Sovereign or Bad Balance Sheets? Euro Interbank Market Fragmentation and Monetary Policy, 2011-2015," .

Gambetti, P., G. Gauthier, and F. Vrins (2018): *Stochastic recovery rate: impact of pricing measure's choice and financial consequences on single-name products*In M. Mili, R. Samaniego Medina, & F. di Pietro (Eds.) *New Methods in Fixed Income Analysis* (pp. 181–203). Cham: Springer.

Gambetti, P., G. Gauthier, and F. Vrins (2019): "Recovery rates: Uncertainty certainly matters," *Journal of Banking & Finance*, 106, 371–383.

Gambetti, P., F. Roccazzella, and F. Vrins (2022): "Meta-Learning Approaches for Recovery Rate Prediction," *Risks*, 10(6).

Geweke, J. (2005): "Elements of Bayesian Inference," in *Wiley Series in Probability and Statistics*, pp. 21–71. John Wiley & Sons, Inc.

Gieseck, A., and Y. Largent (2016): "The impact of macroeconomic uncertainty on activity in the euro area," *Review of Economics*, 67(1), 25–52.

Gordy, M. B. (2003): "A risk-factor model foundation for ratings-based bank capital rules," *Journal of Financial Intermediation*, 12(3), 199–232.

Gorter, J., J. Jacobs, and J. de Haan (2008): "Taylor Rules for the ECB Using Expectations Data," *The Scandinavian Journal of Economics*, 110(3), 473–488.

Grandi, P. (2019): "Sovereign stress and heterogeneous monetary transmission to bank lending in the euro area," *European Economic Review*, 119, 251–273.

Granger, C. W., and Y. Jeon (2004): "Thick modeling," *Economic Modelling*, 21(2), 323–343.

Granger, C. W. J., and R. Ramanathan (1984): "Improved methods of combining forecasts," *Journal of Forecasting*, 3(2), 197–204.

Green, R. C., and B. Hollifield (1992): "When Will Mean-Variance Efficient Portfolios Be Well Diversified?," *The Journal of Finance*, 47(5), 1785–1809.

Greenwell, B., B. Boehmke, J. Cunningham, and G. Developers (2018): *gbm: Generalized Boosted Regression Models*R package version 2.1.4.

Gregory, J. (2012): *Counterparty credit risk and credit value adjustment: a continuing challenge for global financial markets*. Wiley.

Gross, J., and J. Zahner (2021): "What is on the ECB's mind? Monetary policy before and after the global financial crisis," *Journal of Macroeconomics*, 68, 103292.

Gross, M. (2019): "Estimating GVAR weight matrices," *Spatial Economic Analysis*, 14(2), 219–240.

Gu, S., B. Kelly, and D. Xiu (2020): "Empirical asset pricing via machine learning," *Review of Financial Studies*, 33(5).

Hansen, P. R., A. Lunde, and J. M. Nason (2011): "The Model Confidence Set," *Econometrica*, 79(2), 453–497.

Harrison, D., and D. L. Rubinfeld (1978): "Hedonic housing prices and the demand for clean air," *Journal of Environmental Economics and Management*, 5(1), 81–102.

Hartmann-Wendels, T., P. Miller, and E. Töws (2014): "Loss given default for leasing: Parametric and nonparametric estimations," *Journal of Banking & Finance*, 40, 364–375.

Harvey, D., S. Leybourne, and P. Newbold (1997): "Testing the equality of prediction mean squared errors," *International Journal of Forecasting*, 13(2), 281–291.

Harvey, D., and P. Newbold (2000): "Tests for Multiple Forecast Encompassing," *Journal of Applied Econometrics*, 15(5), 471–482.

Hastie, T., R. Tibshirani, and J. Friedman (2008): "Linear Methods for Regression," in *The Elements of Statistical Learning*, pp. 43–99. Springer New York.

——— (2009): *The Elements of Statistical Learning*. Springer New York.

Hofner, B., and T. Hothorn (2017): *stabs: Stability Selection with Error Control*R package version 0.6-3.

Hornik, K., M. Stinchcombe, and H. White (1989): "Multilayer feedforward networks are universal approximators," *Neural Networks*, 2(5), 359–366.

Horvath, R., J. Kotlebova, and M. Siranova (2018): "Interest rate pass-through in the euro area: Financial fragmentation, balance sheet policies and negative rates," *Journal of Financial Stability*, 36, 12–21.

Hothorn, T., K. Hornik, and A. Zeileis (2006): "Unbiased Recursive Partitioning: A Conditional Inference Framework," *Journal of Computational and Graphical Statistics*, 15(3), 651–674.

House, C. L., C. Proebsting, and L. L. Tesar (2020): "Austerity in the aftermath of the great recession," *Journal of Monetary Economics*, 115, 37–63.

Hurlin, C., J. Leymarie, and A. Patin (2018): "Loss functions for Loss Given Default model comparison," *European Journal of Operational Research*, 268(1), 348–360.

Hyndman, R. J., and Y. Fan (1996): "Sample Quantiles in Statistical Packages," *The American Statistician*, 50(4), 361–365.

IMF (2022a): "Global Financial Stability Report: Financial stability risks have risen as war tests the resilience of the financial system through various channels," .

——— (2022b): "Regional Economic Outlook for Europe," .

International Monetary Fund (2021): "Global Financial Stability Report: Preempting a Legacy of Vulnerabilities," .

Jagannathan, R., and T. Ma (2003): "Risk Reduction in Large Portfolios: Why Imposing the Wrong Constraints Helps," *The Journal of Finance*, 58(4), 1651–1683.

Jankowitsch, R., F. Nagler, and M. G. Subrahmanyam (2014): "The determinants of recovery rates in the US corporate bond market," *Journal of Financial Economics*, 114(1), 155 – 177.

Joslin, S., K. J. Singleton, and H. Zhu (2011): "A new perspective on Gaussian dynamic term structure models," *The Review of Financial Studies*, 24(3), 926–970.

Jurado, K., S. Ludvigson, and S. Ng (2015a): "Measuring uncertainty," *American Economic Review*, 105(3), 1177–1216.

Jurado, K., S. C. Ludvigson, and S. Ng (2015b): "Measuring Uncertainty," *American Economic Review*, 105(3), 1177–1216.

Kalotay, E. A., and E. I. Altman (2016): "Intertemporal Forecasts of Defaulted Bond Recoveries and Portfolio Losses," *Review of Finance*, 21(1), 433–463.

Karatzoglou, A., A. Smola, K. Hornik, and A. Zeileis (2004): "kernlab – An S4 Package for Kernel Methods in R," *Journal of Statistical Software*, 11(9), 1–20.

Kilian, L. (1998): "Small-sample confidence intervals for impulse response functions," *Review of Economics and Statistics*, 80(2), 218–230.

Koch, G. R. (2015): "Siamese Neural Networks for One-Shot Image Recognition," .

Kose, M. A., and M. Terrones (2012): "How does uncertainty affect economic performance?," *IMF World Economic Outlook*, pp. 49–53.

Kudô, A. (1963): "A multivariate analogue of the one-sided test," *Biometrika*, 50(3-4), 403–418.

Kuhn, M. (2008): "Building Predictive Models in R Using the caret Package," *Journal of Statistical Software, Articles*, 28(5), 1–26.

Kuhn, M., and K. Johnson (2013): *Applied Predictive Modeling*, New York. Springer.

Kuhn, M., and R. Quinlan (2018): *Cubist: Rule- And Instance-Based Regression Modeling*R package version 0.2.2.

Ledoit, O., and M. Wolf (2003): "Improved estimation of the covariance matrix of stock returns with an application to portfolio selection," *Journal of Empirical Finance*, 10(5), 603–621.

——— (2004a): "Honey, I shrunk the sample covariance matrix," *The Journal of Portfolio Management*, 30(4), 110–119.

——— (2004b): "A well-conditioned estimator for large-dimensional covariance matrices," *Journal of Multivariate Analysis*, 88(2), 365–411.

——— (2012): "Nonlinear shrinkage estimation of large-dimensional covariance matrices," *The Annals of Statistics*, 40(2), 1024–1060.

——— (2017): "Nonlinear Shrinkage of the Covariance Matrix for Portfolio Selection: Markowitz Meets Goldilocks," *Review of Financial Studies*, 30(12), 4349–4388.

Lee, J. D., D. L. Sun, Y. Sun, and J. E. Taylor (2016): "Exact post-selection inference, with application to the lasso," *The Annals of Statistics*, 44(3).

Leow, M., C. Mues, and L. Thomas (2014): "The economy and loss given default: evidence from two UK retail lending data sets," *Journal of the Operational Research Society*, 65(3), 363–375.

Lessmann, S., B. Baesens, H.-V. Seow, and L. C. Thomas (2015): "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research," *European Journal of Operational Research*, 247(1), 124–136.

Liaw, A., and M. Wiener (2002): "Classification and Regression by randomForest," *R News*, 2(3), 18–22.

Liu, J., S. Zhang, and H. Fan (2022): "A two-stage hybrid credit risk prediction model based on XGBoost and graph-based deep neural network," *Expert Systems with Applications*, 195, 116624.

Livshits, I. (2015): "Recent Developments in Consumer Credit and Default Literature," *Journal of Economic Surveys*, 29(4), 594–613.

Loterman, G., I. Brown, D. Martens, C. Mues, and B. Baesens (2012): "Benchmarking regression algorithms for loss given default modeling," *International Journal of Forecasting*, 28(1), 161–170.

Lovell, M. C. (1963): "Seasonal Adjustment of Economic Time Series and Multiple Regression Analysis," *Journal of the American Statistical Association*, 58(304), 993–1010.

Ludvigson, S., S. Ma, and S. Ng (2015): "Uncertainty and Business Cycles: Exogenous Impulse or Endogenous Response?," *NBER Working Papers*, (21803).

Lumley, T. (2017): *leaps: Regression Subset Selection*R package version 3.0.

Lütkepohl, H. (2007): *New Introduction to Multiple Time Series Analysis*. Springer Berlin Heidelberg.

Macchiarelli, C., and P. Koutroumpis (2016): "The current state of financial (dis) integration in the euro area: in-depth analysis," .

Maheu, J. M., and S. Gordon (2008): "Learning, forecasting and structural breaks," *Journal of Applied Econometrics*, 23(5), 553–583.

Makridakis, S. (1986): "The art and science of forecasting an assessment and future directions," *International Journal of Forecasting*, 2(1), 15–39.

Makridakis, S., E. Spiliotis, and V. Assimakopoulos (2020): "The M4 Competition: 100,000 time series and 61 forecasting methods," *International Journal of Forecasting*, 36(1), 54–74.

Mammen, E. (1993): "Bootstrap and wild bootstrap for high dimensional linear models," *The Annals of Statistics*, pp. 255–285.

Mankiw, N. G., R. Reis, and J. Wolfers (2004): "Disagreement about Inflation Expectations," in *NBER Macroeconomics Annual 2003, Volume 18*, pp. 209–270. National Bureau of Economic Research, Inc.

Manzan, S. (2011): "Differential Interpretation in the Survey of Professional Forecasters," *Journal of Money, Credit and Banking*, 43(5), 993–1017.

Markowitz, H. (1952): "Portfolio Selection," *The Journal of Finance*, 7(1), 77–91.

McCracken, M. W., and S. Ng (2015): "FRED-MD: a monthly database for macroeconomic research," Working Paper 2015-12.

McCracken, M. W., and S. Ng (2016): "FRED-MD: A Monthly Database for Macroeconomic Research," *Journal of Business & Economic Statistics*, 34(4), 574–589.

Meinshausen, N. (2017): *quantregForest: Quantile Regression Forests*R package version 1.3-7.

Meinshausen, N., and P. Bühlmann (2010): "Stability selection," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4), 417–473.

Merton, R. C. (1974): "On the Pricing of Corporate Debt: The Risk Structure of Interest Rates," *The Journal of Finance*, 29(2), 449.

Metiu, N. (2012): "Sovereign risk contagion in the Eurozone," *Economics Letters*, 117(1), 35–38.

Michaud, R. (1989): "The Markowitz Optimization Enigma: Is 'Optimized' Optimal?," *Financial Analysts Journal*, 45, 31–42.

Milborrow, S. (2018): *earth: Multivariate Adaptive Regression Splines*R package version 4.6.3.

Miranda-Agrippino, S., and T. Nenova (2022): "A tale of two global monetary policies," *Journal of International Economics*, p. 103606.

Miranda-Agrippino, S., and H. Rey (2020): "U.S. Monetary Policy and the Global Financial Cycle," *The Review of Economic Studies*, 87(6), 2754–2776.

Mishkin, F. S. (1990): "The Information in the Longer Maturity Term Structure About Future Inflation," *The Quarterly Journal of Economics*, 105(3), 815–828.

——— (1991): "A multi-country study of the information in the shorter maturity term structure about future inflation," *Journal of International Money and Finance*, 10(1), 2–22.

Monfort, A., and J.-P. Renne (2013): "Decomposing euro-area sovereign spreads: credit and liquidity risks," *Review of Finance*, 18(6), 2103–2151.

Mora, N. (2015): "Creditor recovery: the macroeconomic dependence of industry equilibrium," *Journal of Financial Stability*, 18, 172–186.

Nakajima, M., and J.-V. Ríos-Rull (2014): "Credit, Bankruptcy, and Aggregate Fluctuations," Working Paper 20617.

Nazemi, A., F. Baumann, and F. J. Fabozzi (2022): "Intertemporal defaulted bond recoveries prediction via machine learning," *European Journal of Operational Research*, 297(3), 1162–1177.

Nazemi, A., and F. J. Fabozzi (2018): "Macroeconomic variable selection for creditor recovery rates," *Journal of Banking & Finance*, 89, 14–25.

Nazemi, A., K. Heidenreich, and F. J. Fabozzi (2018): "Improving corporate bond recovery rate prediction using multi-factor support vector regressions," *European Journal of Operational Research*, 271(2), 664–675.

Nazemi, A., F. F. Pour, K. Heidenreich, and F. J. Fabozzi (2017): "Fuzzy decision fusion approach for loss-given-default modeling," *European Journal of Operational Research*, 262(2), 780–791.

Nazemi, A., H. Rezazadeh, F. J. Fabozzi, and M. Höchstötter (2022): "Deep learning for modeling the collection rate for third-party buyers," *International Journal of Forecasting*, 38(1), 240–252.

Newey, W. K., and K. D. West (1987): "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55(3), 703–708.

Ning, Y., and H. Liu (2017): "A general theory of hypothesis tests and confidence regions for sparse high dimensional models," *The Annals of Statistics*, 45(1), 158–195.

Park, T., and G. Casella (2008): "The Bayesian Lasso," *Journal of the American Statistical Association*, 103(482), 681–686.

Patton, A. J., and A. Timmermann (2010): "Why do forecasters disagree? Lessons from the term structure of cross-sectional dispersion," *Journal of Monetary Economics*, 57(7), 803–820.

Pesaran, M., and Y. Shin (1998): "Generalized impulse response analysis in linear multivariate models," *Economics Letters*, 58(1), 17–29.

Pesaran, M. H. (2021): "General diagnostic tests for cross-sectional dependence in panels," *Empirical Economics*, 60, 13–50.

Pesaran, M. H., D. Pettenuzzo, and A. Timmermann (2006): "Forecasting Time Series Subject to Multiple Structural Breaks," *The Review of Economic Studies*, 73(4), 1057–1084.

Pesaran, M. H., and A. Pick (2011): "Forecast Combination Across Estimation Windows," *Journal of Business & Economic Statistics*, 29(2), 307–318.

Pesaran, M. H., A. Pick, and M. Pranovich (2013): "Optimal forecasts in the presence of structural breaks," *Journal of Econometrics*, 177(2), 134–152.

Pesaran, M. H., T. Schuermann, and S. M. Weiner (2004): "Modeling regional interdependencies using a global error-correcting macroeconometric model," *Journal of Business & Economic Statistics*, 22(2), 129–162.

Pesaran, M. H., Y. Shin, and R. P. Smith (1999): "Pooled Mean Group Estimation of Dynamic Heterogeneous Panels," *Journal of the American Statistical Association*, 94(446), 621–634.

Petropoulos, F., D. Apiletti, V. Assimakopoulos, M. Z. Babai, D. K. Barrow, S. B. Taieb, C. Bergmeir, R. J. Bessa, J. Bijak, J. E. Boylan, J. Browell, C. Carnevale, J. L. Castle, P. Cirillo, M. P. Clements, C. Cordeiro, F. L. C. Oliveira, S. D. Baets, A. Dokumentov, J. Ellison, P. Fiszeder, P. H. Franses, D. T. Frazier, M. Gilliland, M. S. Gönül, P. Goodwin, L. Grossi, Y. Grushka-Cockayne, M. Guidolin, M. Guidolin, U. Gunter, X. Guo, R. Guseo, N. Harvey, D. F. Hendry, R. Hollyman, T. Januschowski, J. Jeon, V. R. R. Jose, Y. Kang, A. B. Koehler, S. Kolassa, N. Kourentzes, S. Leva, F. Li, K. Litsiou, S. Makridakis, G. M. Martin, A. B. Martinez, S. Meeran, T. Modis, K. Nikolopoulos, D. Önkal, A. Paccagnini, A. Panagiotelis, I. Panapakidis, J. M. Pavía, M. Pedio, D. J. Pedregal, P. Pinson, P. Ramos, D. E. Rapach, J. J. Reade, B. Rostami-Tabar, M. Rubaszek, G. Sermpinis, H. L. Shang, E. Spiliotis, A. A. Syntetos, P. D. Talagala, T. S. Talagala, L. Tashman, D. Thomakos, T. Thorarinsdottir, E. Todini, J. R. T. Arenas, X. Wang, R. L. Winkler, A. Yusupova, and F. Ziel (2022): "Forecasting: theory and practice," *International Journal of Forecasting*, 38(3), 705–871.

Probst, P., and A.-L. Boulesteix (2017): "To Tune or Not to Tune the Number of Trees in Random Forest," *J. Mach. Learn. Res.*, 18(1), 6673–6690.

Pykthin, M. (2003): "Unexpected recovery risk," *Risk*, 16(1), 74–78.

Qi, M., and X. Zhao (2011): "Comparison of modeling methods for loss given default," *Journal of Banking & Finance*, 35(11), 2842–2855.

Quinlan, J. R. (1993): "Combining Instance-based and Model-based Learning," in *Proceedings of the Tenth International Conference on International Conference on Machine Learning*, ICML'93, pp. 236–243, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Quinlan, R. (1992): "Learning With Continuous Classes," *Proceedings of the 5th Australian joint conference on artificial intelligence*.

R Core Team (2017): *R: A Language and Environment for Statistical Computing*R Foundation for Statistical Computing, Vienna, Austria.

Ravi, S., and H. Larochelle (2017): "Optimization as a Model for Few-Shot Learning," in *ICLR*.

Ripley, B. D. (1996): *Pattern Recognition and Neural Networks*. Cambridge University Press.

Ripley, B. D., and W. N. Venables (2016): *nnet: Feed-Forward Neural Networks and Multinomial Log-Linear Models*R package version 7.3-12.

Roccazzella, F., and B. Candelon (2022): "Should we care about ECB inflation expectations?," 49.

Roccazzella, F., P. Gambetti, and F. Vrins (2022a): "Correction to: Optimal and robust combination of forecasts via constrained optimization and shrinkage," *International Journal of Forecasting*, 38(3), 1050.

Roccazzella, F., P. Gambetti, and F. Vrins (2022b): "Optimal and robust combination of forecasts via constrained optimization and shrinkage," *International Journal of Forecasting*, 38(1), 97–116.

Rogoff, K. (1985): "The Optimal Degree of Commitment to an Intermediate Monetary Target," *Quarterly Journal of Economics*, 100, 1169–1189.

Saikkonen, P., and H. Lütkepohl (2000): "Testing for the Cointegrating Rank of a VAR Process with Structural Shifts," *Journal of Business & Economic Statistics*, 18(4), 451.

Samuels, J. D., and R. M. Sekkel (2017): "Model Confidence Sets and forecast combination," *International Journal of Forecasting*, 33(1), 48–60.

Santoro, A., S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap (2016): "Meta-Learning with Memory-Augmented Neural Networks," in *Proceedings of The 33rd International Conference on Machine Learning*, ed. by M. F. Balcan, and K. Q. Weinberger, vol. 48 of *Proceedings of Machine Learning Research*, pp. 1842–1850, New York, New York, USA. PMLR.

Santos, A. B., A. A. d. Araujo, J. A. dos Santos, W. R. Schwartz, and D. Menotti (2017): "Combination techniques for hyperspectral image interpretation," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 3648–3651.

Schmitt, M. (2022): "Deep Learning vs. Gradient Boosting: Benchmarking state-of-the-art machine learning algorithms for credit scoring," .

Schuermann, T. (2004): "What do we know about loss given default?," *Wharton Financial Institutions Center Working Paper*, Working Paper No. 04-01.

Shleifer, A., and R. Vishny (1992): "Liquidation values and debt capacity: a market equilibrium approach," *The Journal of Finance*, 47(4), 1343–1366.

Silvapulle, M., and P. Sen (2004): *Constrained Statistical Inference: Inequality, Order, and Shape Restrictions*. Wiley.

Simon, H. (1957): *Models of Man: Social and Rational; Mathematical Essays on Rational Human Behavior in Society Setting*, Continuity in administrative science. Wiley.

Sims, C. A. (1980): "Macroeconomics and Reality," *Econometrica*, 48(1), 1–48.

——— (1986): "Are forecasting models usable for policy analysis?," *Quarterly Review*, 10(Win), 2–16.

Smith, J., and K. F. Wallis (2009): "A Simple Explanation of the Forecast Combination Puzzle," *Oxford Bulletin of Economics and Statistics*, 71(3), 331–355.

Smithson, M., and J. Verkuilen (2006): "A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables," *Psychological Methods*, 11, 54–71.

Soenen, N., and R. Vander Vennet (2022): "ECB monetary policy and bank default risk," *Journal of International Money and Finance*, 122, 102571.

Steinkamp, S., and F. Westermann (2014): "The role of creditor seniority in Europe's sovereign debt crisis," *Economic Policy*, 29(79), 495–552.

Stock, J. H., and M. W. Watson (1999): "Forecasting inflation," *Journal of Monetary Economics*, 44(2), 293–335.

Stock, J. H., and M. W. Watson (2003): "Forecasting Output and Inflation: The Role of Asset Prices," *Journal of Economic Literature*, 41(3), 788–829.

Svensson, L. E. (1997): "Inflation forecast targeting: Implementing and monitoring inflation targets," *European Economic Review*, 41(6), 1111–1146.

Svensson, L. E. O. (2005): "Monetary policy with judgment: Forecast targeting," *International Journal of Central Banking*, (1), 1–54.

Taylor, J. W. (2020): "Forecast combinations for value at risk and expected shortfall," *International Journal of Forecasting*, 36(2), 428–441.

Therneau, T., B. Atkinson, and B. Ripley (2017): *rpart: Recursive Partitioning and Regression Trees*R package version 4.1-11.

Thomas, L., A. Matuszyk, and A. Moore (2012): "Comparing debt characteristics and LGD models for different collections policies," *International Journal of Forecasting*, 28(1), 196–203.

Tibshirani, R. (1996): "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288.

Timmermann, A. (2006): "Chapter 4 Forecast Combinations," in *Handbook of Economic Forecasting*, pp. 135–196. Elsevier.

Tipping, M. E. (2001): "Sparse Bayesian Learning and the Relevance Vector Machine," *Journal of Machine Learning Research*, 1, 211–244.

Tobback, E., D. Martens, T. V. Gestel, and B. Baesens (2014): "Forecasting Loss Given Default models: impact of account characteristics and the macroeconomic state," *Journal of the Operational Research Society*, 65(3), 376–392.

Tu, J., and G. Zhou (2011): "Markowitz meets Talmud: A combination of sophisticated and naive diversification strategies," *Journal of Financial Economics*, 99, 204–215.

Vapnik, V. N. (1995): *The Nature of Statistical Learning Theory*. Springer-Verlag.

Varoquaux, G. (2018): "Cross-validation failure: Small sample sizes lead to large error bars," *NeuroImage*, 180, 68–77.

Vega, S. H., and J. P. Elhorst (2016): "A regional unemployment model simultaneously accounting for serial dynamics, spatial dependence and common factors," *Regional Science and Urban Economics*, 60, 85–95.

Wang, T., R. Liu, and G. Qi (2022): "Multi-classification assessment of bank personal credit risk based on multi-source information fusion," *Expert Systems with Applications*, 191, 116236.

Wang, Z. (2018): *bst: Gradient Boosting*R package version 0.3-15.

Watkins, G. P., and F. H. Knight (1922): "Knight's Risk, Uncertainty and Profit," *The Quarterly Journal of Economics*, 36(4), 682.

Westfall, P., and S. Young (1993): *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*, Wiley Series in Probability and Statistics. Wiley.

Williams, C. K. I., and C. E. Rasmussen (1996): "Gaussian Processes for Regression," in *Advances in Neural Information Processing Systems 8*, pp. 514–520. MIT press.

Winkler, R. L., and S. Makridakis (1983): "The Combination of Forecasts," *Journal of the Royal Statistical Society. Series A (General)*, 146(2), 150.

Wolpert, D. H. (1992): "Stacked generalization," *Neural Networks*, 5(2), 241–259.

Wu, J. C., and F. D. Xia (2020): "Negative interest rate policy and the yield curve," *Journal of Applied Econometrics*, 35(6), 653–672.

Yan, Y., and F. Lin (2016): *FinCovRegularization: Covariance Matrix Estimation and Regularization for Finance*R package version 1.1.0.

Yao, X., J. Crook, and G. Andreeva (2015): "Support vector regression for loss given default modelling," *European Journal of Operational Research*, 240(2), 528–538.

Yao, X., J. Crook, and G. Andreeva (2017): "Enhancing two-stage modelling methodology for loss given default with support vector machines," *European Journal of Operational Research*, 263(2), 679–689.

Yu, M., V. Gupta, and M. Kolar (2019): "Constrained High Dimensional Statistical Inference," .

Zaghini, A. (2016): "Fragmentation and heterogeneity in the euro-area corporate bond market: Back to normal?," *Journal of Financial Stability*, 23, 51–61.

Zarnowitz, V., and L. Lambros (1987): "Consensus and uncertainty in economic prediction," *Journal of Political Economy*, 95(3), 591–621.

Zhang, J., and L. C. Thomas (2012): "Comparisons of linear regression and survival analysis using single and mixture distributions approaches in modelling LGD," *International Journal of Forecasting*, 28(1), 204–215.

Zhao, P., and B. Yu (2006): "On Model Selection Consistency of Lasso," *Journal of Machine Learning Research*, 7, 2541—2563.