Peng Yaohao

Big Data, machine learning and challenges of high dimensionality in financial administration

Brasília - DF

August/2019

Peng Yaohao
Big Data, machine learning and challenges of high dimensionality in financial administration/ Peng Yaohao. – Brasília - DF, August/2019-
240 p. : il. (some colored) ; 30 cm.

Advisor: Pedro Henrique Melo Albuquerque

Thesis (Ph.D.) – University of Brasilia
School of Economics, Business and Accounting (FACE)
Graduate Program in Business Administration
Finance and Quantitative Methods , August/2019.

1. Machine learning. 2. Complexity. 3. Nonlinearity. 4. Bias-Variance Dilemma. 5. Finance. I.Advisor: Pedro Henrique Melo Albuquerque II. University of Brasilia. III. School of Economics, Business and Accounting (FACE). IV. Big Data, machine learning and challenges of high dimensionality in financial administration

CDU 02:141:005.7

Peng Yaohao

Big Data, machine learning and challenges of high dimensionality in financial administration

Thesis submitted to the Graduate Program in Business Administration at University of Brasilia as partial fulfillment of the requirements for attainment of Ph.D. degree in Business Administration, with major in Finance and Quantitative Methods.

The examining committee, as identified below, approves this thesis:

Dr. Pedro Henrique Melo Albuquerque
Advisor

Dr. Herbert Kimura
University of Brasilia

Dr. André Luiz Fernandes Cançado
University of Brasilia

Dr. Flavio Luiz de Moraes Barboza
Federal University of Uberlândia

Brasília - DF
August/2019

Canção da tese

I

Refração
Auroras difusas pontuais

Dispersão
Num âmbar latente, mil cristais
Criam teu saldão

Sísifo: depois desse morro, outro jaz
Vales e cristas, planícies jamais
Se acomode e veja seu Império ao chão

Inspiração
Não parece ser, faz o parecer
Erre nos erros dos errantes banais
Promove na frente que eu arrumo atrás

Interjeição
Posso ignorar, consigo provar
Reverência às referências germinais
Exploro serras com grelhas de pombais

Por diversão
Legado do rei, o porquê já sei
Haltere de chumbo e o peso que traz
Ânfora de corvos, âncora no cais

Pois sim, pois não
Não venci ali, convenci aqui
No próprio opróbrio, vanguarda tanto faz
Em vácuos de plasma, outorga-se a paz

II

Sem desbrio
Tricordiano vil
Ao centro, bela vista
Imponente rente ao frio

Fé candil
Mirando entre os dipolos, sem desvio
Ser profundo, ou simplista
À mercê do juízo ardil

Ervilhei, sigilo enfim
Melhor assim
Conecto o mundo no meu posto de marfim
Relíquias do Medievo, notas secas de um confuso bandolim

Quem sentiu
Turvos pulsos de um caudaloso rio

Prudência, sapiência
Mesclam num ósculo sombrio

Concreto no quadril
Ligo os pontos desse Cosmos baldio

III

Manga com jiló, açafrão e mel
Compilando essa idiossincrasia de sabor
Vem do Leste, faça o teste, sinestésico fervor

Mar dentro do mar, céu além do céu
Vagando eternos fractais buscando o atrator
Ortodoxo, paradoxo, doce e gélido calor

Cinco elementos de construção
Vinte léguas de revelação
Sete emoções de desilusão
Duas dúzias de maturação

Tubo de visões, hipocampo ao léu
Venda seus segundos por sistemas em torpor
Bota a pilha, pega a fila, bate as asas do condor

Álcool em gel
Reflexão em capacetes rotineiros
Mão cruel
Presunção traz ignomínia sem pudor

Ode ao réu
Sem contagem de dez números inteiros
Fogaréu
Foque só no seu histórico clamor

Dono do hotel
Imprimi suas glórias de vapor
Não põe fogo no chapéu
Resta retribuir o favor
Caminhe sem temor
Rumor, louvor
Agora entendo o seu valor

Acknowledgements

Finishing a Ph.D. thesis is a huge challenge, I had the help of many many people throughout the journey, so I'll try to dedicate some words of gratitude to everyone that contributed to this path. I'll do my best as to not leave anyone out, but as there are so many names, I think it's kind of inevitable to end up being unfair to some. Please don't be mad at me if you fail to find your name in this section, be sure that your actions also helped to define me as I am today.

First, I would like to thank my advisor, Prof. Pedro Henrique Melo Albuquerque, for all the assistance and encouragements that goes back to the classes at the undergraduate level. Thank you for being a mentor that taught me not only about mathematical models and scientific research, but also served as a model of a scholar, a teacher, and a friend. Thank you for believing in me from the start, all along you motivated me to always do my best, and still do.

I thank the members of the examining committees at the theoretical essay evaluation and in the qualifying exam – Prof. Thiago Veiga Marzagão, Prof. Patrick Franco Alves, Prof. Flávio Luiz de Moraes Barboza, Prof. André Luiz Fernandes Cançado and Prof. Herbert Kimura – whose comments helped to improve the quality of this work in all stages.

I would like to thank the following Professors from the Graduate Program in Management at the University of Brasilia (PPGA/UnB) for the teachings, advises and encouragements: Prof. Herbert Kimura, Prof. Ivan Ricardo Gartner, Prof. Vinícius Amorim Sobreiro, Prof. Otávio Ribeiro de Medeiros, Prof. Tomas de Aquino Guimarães, Prof. Edgar Reyes Junior, Prof. Francisco Antonio Coelho Junior, Prof. Gisela Demo Fiúza, Prof. Valmir Emil Hoffmann, Prof. Ricardo Corrêa Gomes, Prof. Carlos Denner dos Santos Júnior. I also thank Prof. Daniel Oliveira Cajueiro, whose classes of Finance Theory and Computational Methods I attended and yielded precious learning.

I would like to thank my parents for the support and the understanding along all these years, although not always retributed in the same way. I thank my mother Rao E for having taught me about values, about right and wrong, about the importance of hardworking; I thank you for teaching me Chinese and for telling me that things were badly done when they were indeed badly done: all that's good in me started from you. I thank my father Peng Bo for being a constant mirror for my actions, for showing me the limits of the human civilization, for teaching me about friendship and reciprocity, about "thinking big" and to always look at the whole board: you made me understand the importance of being responsible for all of my actions.

I would like to thank my grandpa Rao Yupu (*in memoriam*), the man known as

"encyclopedia", my personification of wisdom and compassion, I really would like to have known you better. I thank my grandma E Huanqin (*in memoriam*) for the beautiful moments (and the equally important not-so-beautiful moments) back in Beijing; I really miss you, there are many things that I never had the chance to ask you personally, it wouldn't be easier to live with certain answers, but I do keep in me the better side. I thank my other grandma Bai Fengying (*in memoriam*) from whom I inherited my name; I hope to understand better your legacy and carry on some of the values you lived with.

I would like to thank all other family members: my uncles Rao Gang, Rao Lei, Peng Futang (*in memoriam*), Peng Suqin, Zheng Ying, Liu Yihua, Ma Zhongqun; my cousins Rao Keyu (one of my first academic inspirations), Peng Yunbo, Peng Yunhe, Ma Wei, Sophie Rao and Irisa Rao. I've been fairly absent over the decades, but always present in a way as well; the reciprocal also applies.

For my brothers beyond bloodlines (still alphabetically to avoid quarrels...): Afonso Salustiano Neri, Christian Maciel Machado Rocha, João Lucas Magalini Zago, Matheus Gonçalves de Sousa, and Sérgio Thadeu Tavares da Silva Júnior. All along you stood by my side, I'll always be there for you knowing you'll all be there for me should I need it; be sure that I'm the very same person you all once knew (and accordingly mocked on a regular basis...), and the core features that define me will never change. You're part of everything that I have, and your friendships will always be one of my top achievements in this life.

I would express my gratitude to all Professors I had during my academic trajectory at University of Brasilia: Prof. Danielly Silva Ramos Becard, Prof. Carlos Augusto Mello Machado, Prof. Haydée Glória Cruz Caruso, Prof. Virgílio Caixeta Arraes, Prof. Guilherme José da Silva e Sá, Prof. Maria Eduarda Tannuri-Pianto, Prof. Nilton Moura Barroso Neto, Prof. Claudio Ladeira de Oliveira, Prof. Érica Cirino Rosa, Prof. Luciano Martins Costa Póvoa, Prof. Cléria Botêlho da Costa (*in memoriam*), Prof. José Angelo Belloni, Prof. Daniel Oliveira Cajueiro, Prof. Paola Novaes Ramos, Prof. Pablo Holmes Chaves, Prof. Frederico Hartmann de Souza, Prof. Pio Penna Filho, Prof. Ráderson Rodrigues da Silva, Prof. Eduardo Monteiro de Castro Gomes, Prof. Roberto Goulart Menezes, Prof. Luiz Daniel Jatobá França, Prof. Guilherme Scotti Rodrigues, Prof. José Eustáquio Ribeiro Vieira Filho, Prof. André Luiz Marques Serrano, Prof. Pedro Henrique Melo Albuquerque, Prof. Tânia Maria Pechir Gomes Manzur, Prof. Maria Camila Baigorri, Prof. Natália Medina Araújo, Prof. George Rodrigo Bandeira Galindo, Prof. Ulysses Tavares Teixeira, Prof. Ana Flávia Barros-Platiau, Prof. Rafael Tavares Schleicher, Prof. Antônio Carlos Moraes Lessa, Prof. Marcelo de Oliveira Torres, Prof. Carlos Rosano Peña, Prof. Cristina Acciarri, Prof. Maria Helena de Castro Santos, Prof. Ivan Ricardo Gartner, Prof. Matías Alejandro Franchini, Prof. Ricardo Silva Azevedo Araújo, Prof. Victor Gomes e Silva, Prof. Eiiti Sato, Prof. Cristina Yumie Aoki Inoue,

Prof. Carlos Roberto Pio da Costa Filho, Prof. Aline Gomes da Silva Pinto, Prof. Yuri Sampaio Maluf, Prof. Theo Allan Darn Zapata, Prof. Paulo Augusto Pettenuzzo de Britto, Prof. Carlos André de Melo Alves, Prof. Ana Carolina Pereira Zoghbi.

Similarly, I would like to thank every teacher I had during my whole trajectory as a student in elementary and high schools (Escola Classe 411 Norte (2000–2001), Escola Parque 210 Norte (2001), Escola Classe 06 do Guará (2001–2002), and Colégio Militar Dom Pedro II (2003–2010)) and prep school (Exatas, 2008–2010): Carmem, Cinthia, Edilene, Luci, Dalila, Célia, Mara; Rosa, Sandra, Maria Lúcia, Eliane; Lourdes, Edinéia, Érica, Sérgio, Edilena, Alessandra, Jorge, Ribeiro; Gercilo Alves, Paulo César, Bonfim, Maciel, Ronnie, Paz, Inácio, Luciano Macedo, Krüger, Hélio Veloso, Sergio Tomazelo, Dargon Afonso, Fortuna, Dalveci, Emerson Freitas, Cleandro Barbosa, Lobato, Roberlândio, Santiago, Tocantins, Cardoso Filho, H. Santana, Antônio Da Silva, Fábio, Paulo César, César Quintanilha, Roberto Sangaleti, Edson Barroso, Márcio Matos, Júlio César Faria, Athos, Márcio Dantas, Zanina, Azevedo, Alckmin, Hermínio, Eulino, Paulo Costa (*in memoriam*), P. Sousa, S. Sousa, G. Santana, P. Santana, Hernandez, Victor, Elaine, Gildo, Wernes, Michel, Milton Gomes, J. Borges, Fábia, Aprígio, João Gonçalves, Edson Mesquita, Raphael, Fernanda, Edson Márcio, Luciano, Goulart, Gontijo, Monteiro, Sócrates, Lera, Ademésio, André Assis, Paulo Domingues; Ronei Castro; Otávio, Poliana; Viviane, Márcia Bomfim, Valéria Teixeira, Eliana Aquino, Rafaela, Luise Martins, Ana Paula, Érica Corrêa, Zilda Vilarins, Cassiano Barra, Fábio, Silvia Tominaga, Geovanildo; Leiliane, Orlando Desidério, Réder, Rodnei, Benilva, Adriano Vieira, Almeida, Edeilson Cavalcante, Edgard Cândido, Diogo Pelaes, Ronan Amorim, Aclésio Moreira; Arineide, João Emílio Moreira, Marcelo Henrique, Rejane Morena, Kélvia Lustosa, Evandro Avocado, Rêiner Godoy, Jone Borges; Angélica Gislene, Jeferson Maximino, Ruan, Leonardo Mendes, Paulo Rufino, Antônio Nilmar Nascimento, Michelle Cristina; Afonso Paz, Marcos Laurindo, Luiz Paulo, Marcelo Aguiar, Cléber Alves, Henrique, Gilberto Marques, Leonardo David, Jairo Soares, Elijaime, Wescley Oliveira, Verônica, Manoel Everton, Isaac Lacerda, Ronaldo Rodrigues, Elisângela Aparecida, Luciana Perdigão, Mike; Gersoneide, Érika Hitzschky, Caroline Kochenborger, Brunye, Jaqueline, Dalva Costa; Maíra Zenun; Fabiana, Rômulo Borges, Gilmara de Queiros, Vicente Rocha, Viviane Ataíde, Fillipi Oliveira, Rênia Karina; Lindsay Baptista, Cleiciane Lobato, Juliana; Sandra, Elisa, Wanderlei, Rosemary, Sandra, Lucilene, Vera Lúcia, Werton Dias, Cleide Ximenis, Antônio Rivael; Tatiana Trindade, Márcia Lima, Michelle Paiva; Sandra Melo, Juliano Neiva; Jeronymo, Kesley, Karine, Sérgio Guerra, Leonardo Abrantes; Garcia, Antônio Melo, Genivaldo, Caroline, Alessandro Girotto, Jorge Silva, José Júnior; Camilla Osiro, Juliana, Cléa Maduro, Tiago, Silvia Tominaga, Câmara, Viviane Faria, Messias, Luis Batista, Wesley Ferdinando, Pedro Marinho, Hildebrando Alves, Diogo Gomes, Morais, Harudgy Amano, Bruno Mangabeira, Sormani, Ronaldo Sartori, Ailton, Márcio Padilha, Edilberto, Kléber, Washington Cândido, Paulo Macedo, César Severo, Hilton Chaves,

Gustavo Conforto, José Maria, Gastão Sanches, Patrícia, Cleiton Acacio, Cristóvão Resende, Sami, Renato Sato, Guilherme Dias, "Queijo" (never figured out your real name, though), Adailton, Fernando Barbosa, Rick, Cláudio, Antônio Rivael, Fernando Borges, Rômulo, Márcio, Luciana, Paulo.

Every time I am called "Teacher" or "Professor" I tend to think of all the scholars I once called "Teacher" or "Professor"; these are indeed very special titles that carry a very special responsibility – the implication of having to teach knowledge and tools that ultimately reshapes the life of the student. I kept teachings and experiences from every Professor I ever had, always improving over the good experiences and being cautious about the not so good ones. Those who learn from the past can build a better future, and I am always trying to learn from my past mentors. There's a traditional saying in China that states: "Master for a day, father for a lifetime". Your teachings will remain in me forever, and I'm truly grateful for that.

I would also like to thank many other professors from whom I didn't take classes, but I had conversations with or had access to works, teaching material and experiences which served as references to me in many different ways: Prof. João Carlos Félix Souza, Prof. Celius Antonio Magalhães, Prof. Vânia Carvalho Pinto, Prof. José Guilherme de Lara Resende, Prof. Jhames Matos Sampaio, Prof. Paulo Henrique Pereira da Costa, Prof. Ricardo Ruviaro, Prof. Henrique Costa dos Reis, Prof. Yuri Dumaresq Sobral, Prof. Pedro Henrique Verano Cordeiro da Silva, Prof. Patrícia Santos, Prof. Joaquim Jose Guilherme de Aragão, Prof. Ary Vasconcelos Medino, Prof. Donald Matthew Pianto, Prof. Wilson Toshiro Nakamura, Prof. Kárem Cristina de Sousa Ribeiro, Prof. Alethéia Ferreira da Cruz, Prof. João Mello da Silva, Prof. Jose Renato Haas Ornelas, Prof. Eduardo Ottoboni Brunaldi, Prof. Aldery Silveira Júnior, Prof. Evelyn Seligmann Feitosa, Prof. Caio César de Medeiros Costa, Prof. Paulo Roberto Cardoso, Prof. Siegrid Guillaumon Dechandt, Prof. Evaldo César Cavalcante Rodrigues, Prof. Andrea Felippe Cabello, Prof. Victor Rafael Rezende Celestino, Prof. Nigel John Edward Pitt, Prof. Eduardo Yoshio Nakano, Prof. Cibele Queiroz da-Silva.

In special, my gratitude goes to Prof. Olinda Maria Gomes Lesses, Prof. Marcos Alberto Dantas, and Prof. José Márcio Carvalho, for having granted me various opportunities to teach at the undergraduate level in a top university like the University of Brasilia. The experience I gained from those eight semesters were absolutely priceless and extremely important for my whole academic career.

I would like to thank my fellow researchers from the Machine Learning Laboratory in Finance and Organizations (LAMFO): Sarah Sabino, Cayan Portela, Mariana Montenegro, Igor do Nascimento, Matheus Facure, Stefano Dantas, Lucas Gomes, Pedro Alexandre Barros, Ana Julia Padula, Jáder Martins, João Gabriel Souza, Gustavo Monteiro, João Victor Machado, Fernanda Amorim, Patrick Henrique, Leonardo Galler, Rafael

de Morais, Alfredo Rossi, Victor Matheus, Pedro Correia, Hugo Honda, Marcelo Felix, Maísa Aniceto, Daniel Viegas, Ennio Bastos, Matteo Kimura, Renê Xavier, Paulo Vítor Barros, Fábio Medina, Patrick Tatagiba, Carlos Aragon, Lore Bueno, Patrick Franco, Lívia Batalha, Silvio Gomes, Felipe Natan. Being with you guys is a constant learning experience for me, I am really proud to be part of this elite research group. Having assembled so many talents in one place shows how promising this group is and that we are indeed at the edge of future-building. I hope LAMFO has been an environment in which everyone is free to share and to learn, in every project I get involved I am able to learn lots of cool stuff from every one of you, I really hope that you have been enjoying as much as I have.

I would like to thank my friends and colleagues from the Brazilian Institute of Applied Economic Research (Ipea): Lucas Mation, Guilherme Cassaro, Daniel Viegas, Tamara Vaz, Rachmyne Diabaté, Vinícius Oliveira, Rafael Franco, Nícolas Pinto, Adriano Torres, Gustavo Coelho, Igor Camelo, Iasmini Lima, Isis Moura, Ane Caroline, João Victor Machado, Rodrigo Arruda, Matheus Schmelling, Edgar Sampaio, Luciano Moura, Fernanda Amorim, Adriano Figueiredo Torres, Jáder Martins, Gustavo Monteiro, Lucas Gomes, Carlos Aragon, Renata Dias, Welligtton Cavedo, João Marcello Schubnell, Pedro Garcia, Leonardo Jesus, Pedro Altomar, Pedro Muniz, Talita Lima, Ana Julia Padula, Alixandro Werneck, Jean Sabino, Michel Alba, Lucas Henrique, Sérgio Viriato, Sérgio Coletto; Nilo Saccaro, André Zuvanov, Raimundo da Rocha, Thaynara Martins, João Gabriel Souza, Gabriel Rabello, Isaac Eberhardt, Rennaly Sousa, Luísa Dusi, Bianca Paiva, Jéssica Fernandes, Ludmilla Mattos, Anaely Machado, Cayan Portela, Ernesto Lozardo, Carlos von Doellinger, Leonardo Monastério, Caio Nogueira, Guilherme Duarte, Gustavo Basso, Daniel da Mata, Éder Gillian, Fabíola Vieira, Roberta Vieira, Edvaldo Batista, Alexandre Cunha, Graziela Ansiliero, Marina Garcia, Alexandre Ywata, Adolfo Sachsida, Rogério Boueri, Luis Kubota, Aguinaldo Maciente, Fabiano Pompermayer, Bernardo Schettini, Lucas Vasconcelos, João De Negri, Carlos Corseuil, Vanessa Nadalin, Israel Andrade, Solange Ledi, Danilo Severian, Alex, Lidiane, Neta, Humberto Sousa, Dalva de Aguiar, Ester Antonia, Mariana, Matheus, Marcela Costa, Roberta Ferreira, Adriana Ito, Mylena Fiori, João Cláudio, Antenor Francilino, Lana, Fátima, Amanda, Fernando, Renata, Hélio, Vânia, Lígia, Nathália, Pedro, Flávia, Cecília, Mauro, Robson, Rodrigo, Fernanda, Felipe, Alcide, Josué, Luiz Bahia, Giovanni Roriz, Diogo Chaves. Ipea is one of the main think-tanks of Brazil, there are brilliant minds in every corner of that building, even apparently unpretentious desk conversations can provide unique learning opportunities, not only in a theoretical level but also in making the theory work in the practice. In special, working in the Data Science Lab made me understand more clearly the importance of many invisible efforts involved behind a nice and tidy dataset, as an amazing amount of work is often needed in the back-end so someone can simply push some buttons to yield analysis that ultimately can define the course of action of the

Brazilian government and consequently affect the lives of millions of people.

I would like to thank the secretary staff from various departments from schools and the University of Brasilia, who were helpful when requested: Germano, Edson, Adriana, Leonardo Senise; Jailma, Rodrigo, Karina; Sonária, Selma, Gustavo, Edvânia, Victória, Keila Rossana; Ana, Emília, Elizânia, Taciano, Alexandre, Aldeni, Júlia, Hector, Roseane, Celi, Francine, and Vanderlei. Many thanks to Edimilson Marinho for the support given during the capacitation course in June 2019.

I also thank all professors and researchers involved in ANATEL/IBICT's project: Prof. Carla Peixoto Borges, Prof. Rafael Barreiros Porto, Prof. Eluiza Alberto de Morais Watanabe, Prof. Jorge Mendes de Oliveira-Castro Neto, Prof. Pedro Henrique Melo Albuquerque, Prof. João Carlos Neves de Paiva, Prof. Antônio Isidro da Silva Filho, Lucas Gomes, Guilherme Lopes, Fernanda Amorim, Daniel Viegas, Denise Oliveira, Alcino Franco, Jéssica Barbosa, Érica Perli, Alexandre Gameiro, Fábio Koleski, Rodolfo Angelini, Priscila Reguffe, Andreza Oliveira, Fabiana, Poliana Sepulveda, Daiane Kachuba, Jéssica Bilac, João Batista Machado, Lílian Mazzocante, Giovanna Chita, Tiago Braga, Alexandre Oliveira, João Vitor Coelho, Patrícia Luque, Tatiana Queiroz, Nathaly Rocha. I had a great opportunity to learn from all of you.

I would like to thank all colleagues from the 2011/1 class of International Relations at University of Brasilia, as well as others who joined throughout the other semesters: Rafael Leite, Pedro Henrique Nascimento, Vítor Olivier, Felipe Fortes, Lucas Caldeira, Carlos Solis, Alysson Soares, Augusto Leonel, Roman Leon Neto, Eurides Viana, Mauro Medeiros, Rodrigo Juaçaba, Hugo Luis, Joel Barini, Erick Colina, Caio Vivan, Murillo Feitosa, Márcio Carvalho, Victor de Sá, Vítor Garcia, Vinícius Matsuyama, Ana Carolina Wallier, Laura Fonseca, Mariana Rios, Marina Ventura, Ana Carolina Capeletto, Mariana Macêdo, Raíssa Vitório, Bárbara Bueno, Ricardo Prata, Luisa Helena, Luisa Merico, Renata Lourenço, Mariana Dias, Paula Matheus, Beatriz Bazzi, Nathália Vieira, Manuella Brigitte, Carolina Jordão, Carolina Thaines, Juliana Akemi, Fernanda Fernandes, Larissa Bertolo, Caroline Blasque, Isabela Damasceno, Thábita Pereira, César Carvalho, Juliana Grangeiro, Isidoro Eduardo, Luiz Artur Costa, Ebenésio Ambrósio, Nelson Veras, Christiane Najar, Pedro Batista, Thelma Sodré, Najme Simon, Ludmyla Nellen, Thamires Quinhões, Juliana Cintra, Júlia Rovery, Bárbara Vilarinhos, Cássio Akahoshi, José Carlos Amaral, Paula Kraetzer, Ananda Martins, Graziela Streit, Pedro Melo, Letícia Laurentino, Deborah Ribeiro, Natalia Sardenberg, Joaquim Otávio, Maíra Minatogau, Eduardo Cavalcante, Victoria Veiga, Vitória Moreira, Evellyne Aparecida, Talita Melgaço, João Paulo Nacarate, Lucas Arruda, Safiya Yusuf, Rachel Scott, Leandro Teles, Juliane Menezes, Nathália Borges, Lucas Carvalho. I wasn't much of a talker during my undergraduate studies, but I was a fairly good listener, and I learned much upon hearing and seeing my predecessors and contemporaries. My period as a student

of International Relations certainly allowed me to diversify my analytical tools and social experiences; even though my academic expertises have drastically shifted over the years, I learned much from my "qualitative" years and they surely gave me a interesting background to a whole different way of critical thinking and problem-solving.

I would like to thank my seniors from older semesters of International Relations and colleagues I had scattered across my undergraduate classes at the University of Brasilia and other circumstances (ordinary or odd) throughout the semesters: Mário Frasson, Thaís Oliveira, Patrícia Martuscelli, Jéssika Santiago, Márcio Alves, Deborah de Castro, Ivan Bastos, Clara Aragão, Kairo Aguiar, Rômulo Albernaz, Aloir Martins, Luísa Maria Cavalcante, Cárita Cavalcante, Marcellus Lopes, Renata Emerick, Marcos Valente, José Luís Alves Fernandes (*in memoriam*), Letícia Raimundo, João Sigora, Maíra Carvalho, Stefanos Drakoulakis, Lincoln Guabajara, Mariana Nóbrega, Dominique Paé, Victória Monteiro, Peter Trakofler, Corina Nassif, Nicolas Wulk, Paulo Shinji, Rodrigo Guerra, Tiago Veronesi, Lucas D'Nillo, Marina Calonego, Michael Dantas, Marcelo, Mariana, Heitor, Carlos Góes, Ulysses Bispo, Marcius Lima, Lucas Corá, Gabriel de Magalhães, Felipe Calainho, Marina Uchôa, Natalia Kelday, Priscila Pfaffmann, João Brites, Nicolas Powidayko, Felipe Dias, Jonathan Braga, Débora Lobato, Gabriella Passaglia, Bruno Suzart, Ana Júlia Fernandes, Thayana Tavares, Emanuelle Rocha, Jéssica, Davi Leite, Yuri , Leonardo Bandarra, Gustavo Ziemath, Flávia Said, Caio, Rebeca Machado, Amina Nogueira, Diogo, Leonardo, Ricardo, Jeremy Paule, Tarsis Brito, Sofia Fernandes, Ana Carolina Campos, Augusto Sticca, Igor Albuquerque, Rhebecca Cunha, Luisa Kieling, Ana Luísa Alves, Fernanda Burjack, Eduardo Sena, Leo Tavares, Fabiane Freitas, Eduardo Izycki, Flávia Neme, Vinnie, Heitor Torres, Mateus Reis, Daniel Reis, Juliana P., Flávia Camargo, Bárbara Borges, Renata Carneiro, Andrei Ricardo, Beatriz Félix, Daniel Luzardo, Alan Melo, Judith Aragão, Emília Aragão, Jéssica Silva, Lucas Costa, Deborah Y., Matheus Lamounier, Rodrigo Ferrari, Felipe Fassina, César, Roberta Miranda, Érica Costa, Amanda de Andrade, Amanda Ferreira, Gabrielle Rangel, Thuany, Bárbara, Rafaella, Luciana, Elias Couto, Samuel Leal, Vítor Lima, Izabel Brandão, Ricardo Lobato, Jorge, William, Júlia Pizzi, Humberto Firmino, Rômulo Ataídes, Bruno Godoy, Izabel Guevara. All of you contributed to my academic trajectory in some way, I'll kindly remember all the stories worth remembering that I shared with each one of you.

I thank all teaching assistants and monitors I had during my academic life (unfortunately, too much to mention or to recall everyone, due to the unified disciplines...): Larissa Bertolo, Elisa Dall'Orto, Camilla, Laís, Alessandro, Janete, Louise Cugula, Daniel Cavalcanti, Daniel Dinelli, Eduardo, Jéssica Gonçalves, Clarissa Palos, Paulo, Louise Martins, Fernando Couto, Tamara Vaz, Ludmilla Müller, Diogo, Yasmim Odaguiri, Jean Lima, Giordano de Almeida, Alberto André Francisco, Prof. Fernanda Ledo Marciniuk. I still carry many teaching experiences I had with you in my own lessons, in special concerning the naturally smaller "distance" between a senior undergraduate student and his

juniors. Particularly, the monitors from “Introductory Economics” (IEMonit) were especially solicitous and helped to inspire my interest in teaching (even though I was rejected in the selection process for period 2011/2, and even to this very day I didn’t receive the feedback I was promised...): Jéssika Santiago, Max Villela, Daniel Pina, Lorenna Vieira, João Negreiros, Luis Guilherme Batista, Letícia Raimundo, Nicolas Wulk, Lucas Bispo, Camila Wahrendorff, Haigo Porto, Ana Júlia Monteiro, Débora Jacintho, Luana Drumond, Samantha Vitena, Anna Paula, Thayana Tavares, Renata Motta.

Special thanks to Prof. Pedro Henrique Melo Albuquerque and my colleagues of the class MMQD2 of period 2013/2, this was indeed a “prized package” that gathered some of the best I have ever seen, and this class was the genesis to many elite researchers in machine learning that later constituted the original core of LAMFO: Sarah Sabino, Mariana Montenegro, Pedro Alexandre Barros, Thiago Raymon, Fábio Medina, Ludmilla Müller, Leo Saracan, Daniel Pagotto, Daniel Carvalho, and Lucas Pinheiro. Without exaggerating, this class really changed my academic path, I am forever grateful to all of you.

I would like to thank my friends and colleagues from the Graduate Program in Management at the University of Brasilia (PPGA/UnB) and from the classes I picked from the Economics Department: Thiago Raymon, Leonardo Bosque, Sarah Sabino, Yuri Maluf, Mariana Montenegro, Pedro Correia, João Gabriel Souza, Emmanuel Abreu, Luiz Medeiros, Raphael Pereira, Gustavo Basso, Alexandre Leite, Jorge Barbosa, Monique Azevedo, José Rômulo Vieira, Bruno Miranda, Marina Garcia, Leonardo Magno, Marcelo Godinho, Wanderson Bittencourt, Paulo Sérgio Rosa, Raphael Brocchi, Daniel Tavares, André Porfírio, Nathália Melo, Mário Salimon, Natasha Fogaça, Lana Montezano, Mariana Rêgo, David Bouças, Pablo Pessôa, Alexander Dauzeley, Gustavo Alves, José Nilton, Everton Verga, Sérgio Freitas, Isabela Ferraz, Ana Carolina Costa, Júnia Falqueto, Manoel Fonseca, Silvia Onoyama, Ricardo Ken, Marilú Castro, Juliana Moro, Emília Faria, Oscar Oliveira, Carolina Sgaraboto, Nazareno Marques, Alex Fabiane, Marcelo Cardoso, Rodrigo Montalvão, Vivian Caroline, Ana Paula Lopes, Ladilucy Almond, Leovanir Richter, Luis Fernando Pinto, Paulo Góes, Walter Faiad, Bruno Saboya, Alencar Libâneo, Rafael Farias, Jéssica Traguetto, Hécio Almeida, Hannah Salmen, Bárbara de Medeiros, Matheus Rabetti, Renata Telles, Tiago Rusin, Eduardo Lafetá, Cristiano Lúcio, Leonel Cerqueira, Marcelo Finazzi, Amanda Paiva; André Maranhão, Saulo Benchimol, Cauê Mello, Marcelo Cruz, Guilherme Paiva, Camila Pereira, Prof. Santiago Ravassi, Edmar Rocha, Prof. Mauro Moraes Alves Patrão; Noël Olokodana; Diogo Picco, Vanderson Delapedra, Felipe Vilhena, Laerte Takeuti, Andreia Barros, Iago Cotrim, Ludmilla Mattos, André Veras, Fioravanti Mieto, Denise Oliveira, Érica Botelho, Diogo da Fonseca, Alex Rojas, Douglas München; Patrícia Rosvadoski, Newton Miranda Junior, Dayse Carneiro, Bernardo Buta, Jeovan Assis, Isadora Lopes, Eloisa Torlig, Rommel Resende, Oscar Rocha, João Maria de Oliveira, Alice Plakoudi, Eduardo Rubik, Gabriel de Deus,

Mayra Viana, Eduardo Leite, Luciana Cualheta, Jéssica Fernandes, Danilo Ramos, Bruno Braga, Maricilene do Nascimento, Marilene de Oliveira, Marizaura Camões, Pedro Antero, Clarissa Melo. PPGA was a very friendly environment, the years I lived in there showed me that even potentially fierce rivals could be friends that encourage each other and make each other stronger. I made very good friends during the classes, presentations, and conferences, I could learn from works I had no idea of the basics of the subject. The relationship with my fellow researchers was the best possible, I always felt extremely respected amongst my colleagues, and I also always treated everyone with due respect. I could learn not only from colleagues from finance with a similar way of thinking from me but also many scholars with quite different motivations and *modus operandi*, this certainly made me a more complete academic researcher.

I would like to thank all students I taught in various disciplines at the undergraduate level. Now there are really too many names, so I won't cite everyone, just a few that I recall right now: Maria Gabryella, Lívia Aragão, Letícia Mendonça, Marcela Coutinho, Guilherme Falcão, Guilherme Urbano, Rodrigo Ferreira, Marcelo Alcântara, Natália Amorim, Marina Barros, Pedro Ferreira, Paula Macedo, Salomão Ferretti, Marcelo Babilônia, Cícera Monallysa, Monique Novais, Renata Patriota, Ana Karina, Luca Torres, Lauro Pedreira; Alisson Cavalcante, Laíra Brito, Eduardo Dantas, Eduardo Bogosian, Eduardo Fleury, Rubens Toledo, Átila Gomes, Verônica Mamede, Túlio de Paiva, Túlio Vicente, Marcella Turon, Pedro Michel Sinimbu, Lean Nascimento, Isabela Nepomuceno, Fernanda Pinheiro, Daniel Jatobá, Leandro Santiago, Marcos Mousinho, Vítor Neiva, Victor Luciano, André Santana, Filipe Matos, Fernanda Lima, Bruno Godoy, Matheus Gonçalves, Rodrigo Sampaio, Luiz Felipe Paulino, Luísa Cavalcante, Gustavo Fischer, Yuri Cabral, Samantha Santana, Jacy Rodrigues, Joicy Keilly, Luiz Fernando Soares; Matheus Kempa, Flávio Herval, Igor Martins, Matheus Périco, Gustavo Damasco, Gustavo do Espírito Santo, Bernardo Ravanello, Luisa Bonfá, Rodrigo Matos, Thaís Testoni, Thaís Luiza, Cristina Cavaletti, Cláudia Veloso, Ludmila Rocha, Rebecca Viana, Bárbara Braga, Pedro Paulo França, Camila Melo, Felipe Camargo, Sofia Porto, Abílio Phellipi, Alex Iaccino, Ana Luísa Tomassini, Brenda Giordani, Fernanda Scafe, Matheus Martins, Diego Tolentino, Thiago Amaral, Thiago Lima, Daniel Milazzo, Ivan Mello, João Júlio, Guilherme Scattone; Jordane Reis, Michele Gasparoto, Camilla Zorzi, Ismael Santos, Raíssa Pires, Ludmila Boaventura, Thaís Santos, Álvaro Bragança, Thainá Chavarry, Matheus Corrêa, Claudio Cavalcante, Bruna Martins, Bruna Nunes, João Victor Machado, Felipe da Rocha, Ycleda Oliveira, Pedro Guerra, Khalil Santarém, Vitor Aragão, Luísa Versiani, Lucas Costa, Gustavo Buta, Felipe Henrique Alves, Eglay Moreno, Brian Hebert, Felipe Jardim; Raquel Félix, Jéssika Siqueira, Karina Ferreira, Karine Rangel, Leonardo Serikawa, João Vítor Alencar, Fernanda Amorim, João Vítor Borges, Gabriel Benigno, Danielle Leite, Gabriela Cristina, Pablo Almeida, Carolina Rodrigues Lopes, Alexandre Fantini, Maria Tereza Córdova, Gabriel Homem, David Eloi, Clenilson

Costa, Caroline Raposo, Amanda Pinheiro, Glauber Níkolas, Maria Raquel Vera, Thiago Ramiris, Talyta Soyer, Marina Siqueira, Diego Souza, Caio Júlio César; Tallyrand Jorcelino, Scarlett Rocha, Wilson Santiago, Moisés Có, João Vitor Soares, Helena Cartaxo, Georgia Nasr, Luiza Rosendo, Vítor Lago, Carla Santana, Flávia Santana, Lorranna Couto, Maciel Neri, Evandro Garcia, Carlos Alexandre dos Santos, Déborah Carvalho, Ana Paula Lima, Paulo Braga, Marina Iwakiri, Matheus Marinho, Renan Kuba, Vítor Capistrano, Luiz Pimenta; Bernardo Pereira, Fernanda Gabriele, Albert Dobbin, Maria Júlia Gonçalves, Luis Batista, Elisa Silveira, Taian Cristal, Natalia Lustoza, Natasha Veloso; Divaldo Antonio, Gustavo Andrino, Gabriel Fontes, Pedro Ido, Helena Kichel, Fabiana Mariquito, Catharine Pedrosa, Mariana Galvão, Gabriela Almeida, Cristiano Cardoso, Eduardo Freitas, Daniel Viegas, Davi Resende, Pedro Altomar. I always heard that teaching was the best way to learn, but I never truly understood this until I had to actually teach, after which I found this to be true. It is a true challenge to be at the "other side", all attention of the class focused and a little space for mistake. It has been a wonderful experience ever since the first time I had to teach at the undergraduate level, and it's amazing how a same concept and class can provide the teacher quite different experiences even after years. I already had a lot of students, among which a great number helped me actively to become a better teacher; my gratitude goes to all of you (naturally including those I didn't cite here), I hope you could learn from me as much I have learned from you, all of you provided me with opportunities to improve myself and inspired me to become a better teacher.

My gratitude also goes to all assistants I had for my classes at the undergraduate level: Marcellus Lopes, Noël Olokodana, Gleison Batista, Bárbara Braga, Matheus Raposo, Fernanda Amorim, and Brenda Baumann.

I would like to thank my coauthors in scientific researches for the persistence and patience during our academic partnerships: Jáder Martins, Igor do Nascimento, Mariana Montenegro, Ana Julia Padula, João Victor Machado, Cayan Portela, Rafael de Morais, Leonardo Bosque, Marcelo Félix, Pedro Alexandre Barros, Sarah Sabino, Matheus Facure, Gustavo Monteiro, Fernanda Amorim, Matheus Kempa, Emmanuel de Abreu, Luiz Medeiros. Your contributions will always be very kindly remembered, as they're forever bound into the results that we achieved together. All harshness and hurry aside, putting up with me do have its payoffs...

I thank all students whom I advised in scientific projects and term papers: Henrique Passos, Patrick Tatagiba, Mateus Rodrigues, Matheus Kempa, Marcela Coutinho, Rafael Barros, and André Veras. The advising experience allowed me to be even further on the "other side", to see the learning process from the other perspective and to better understand how to conduct well the aspirations of a motivated student.

I thank for the opportunities I had to evaluate the term papers of the following

students as a member of the examining committee: Marcelo Félix, Matheus Alves, Rômulo Albernaz, Giovanna Rocha, Helberth Macau, Enrico Eduardo, Plínio de Sousa, Larissa Thaís, Pedro Filgueiras, Alexandre Bernardes, and Túlio Cavallini. To judge can appear easy, but to do so fairly can be demanding, I had opportunities to do so some times in my life. I also thank the advisors that invited me to compose the examining committees, as well as the students, who reacted in different ways, making me rethink about my criteria after reviewing each work.

I arrived in Brazil when I was 5, without knowing a single word in Brazilian Portuguese; the first months were really hard, although I probably never truly realized the size of that challenge. I grew in an unknown environment filled with stares of perceived weirdness, sometimes incomprehension, but at most times curiosity. I had some wonderful teachers and colleagues, others not so wonderful, but equally important to my formation, I keep all of them with nostalgia and great consideration. Jokes and immaturities aside, today I'm grateful for it all and I am glad to be able to recall all of it with a laugh. I dearly miss the vast majority of those moments (even the other minority is also really worth to be remembered... actually they are the ones that make you truly understand the value of the good stuff), they helped to define the person I am today. Being joyful or painful, all those moments are definitely worth remembering.

At this point there are just so many names, I certainly forgot some important ones. Still, even those I forgot to list here contributed to my life, please don't mind if you're not here; hopefully, you can let me off the hook...

Let's do this chronologically to make it easier.

2000: Gabriela, Camargo, Reginaldo, Eduardo, Maciel, Germano, Francisco, Sharon, Paloma, Mateus, Érika, Ana Maria, Bruna, Rayanne, Rayana, Lucas, Vítor, Priscila Maria, Rudne, Bianca; Raimundo, Iago, Itamar, Elói, Rico, Jadenilson, Tiago, Renan, Nilton.

2001: Ronaldo, Joana, Douglas, Rafael, Gabriel, Priscilla, Pedro Henrique, Gabriela, Jéssica; Roberto Arnaldo, Arenaldo, Any, Gustavo, Lucas Porto, Lucas França, Eric, Adriano, Robertinelli, Samuel Pimenta, César, Marcos Cazú, Tiago Melo, Thiago Henrique, Breno, Maurício Mettino , Paulo , Max, Raíssa, Jorge Sabino, Anderson Luque.

2002: Wendell, José Nilson, Caio, Danilo, Gabriela Roxanna, Gabriela Corrêa, Rafael, Guilherme, Gisele, Kelly, Diego, Marcos Emmanuel, Christian, Patrícia, Rafaela, Paulo, Horácio, Luciano, Rafael, Ernesto Tudor, Remy Nobre.

2003: Rômulo Paulino, Ítalo Almeida, Thiago Matos, João Vítor Pacheco, Victor, Daniel Noble, Emmanuele, Alessandro, Alexandro, Andressa Holanda, Andressa Pompas, Douglas Alves, Douglas Silva, Higor Santana, Marina, Raíssa Freire, Natália Maria, Nathália Lamosa, Rafaela, Roberta, Taís Martins, Thaís Cabral, Roberta, Wendy, Luiz

Felipe Medeiros, Eike Lobato, André, Gabriel, Hugo, Marcus Vinícius, Paulo, Pâmella Duarte, Kayo Eduardo, Reginaldo Sousa, Bárbara Virgínia, Kallycia Bose, Nathan Yohan, Matheus Carvalho, Diego, Victor Hugo Mee, André Rabelo, João Carlos, Fernando Bento Filho, João Ciriaco, Eustáquio Fortes.

2004: Jaqueline Azevedo, Isabela Cardoso, Jéssica Maria, Douglas Alves, Diógenes, Diane Desirée, Roberto Arnaldo, Thiego Lorran, Tarik El-Harim, João Paulo Nogueira, Ananda Jade, Bruna Macêdo, Hélio Veloso, Natan Rangel, Karolline Brito, Guilherme Marques, Natália, Leandro Bertolazi, Marcos Fabrício, Michael Militão, Micael, Lucas Façanha, Brayan Leandro, Thiago Oliveira, Ícaro, Rebeca Iviê, Débora Azevedo, Stephany, Jorge Yuri, Radígia Mendes, Frederico C. Borges, Luiz Bastos, Kayke, Jean Victor, Felipe Bispo, Samuel, Larissa Melo.

2005: Afonso Neri, Jéssica Ferreira, Thalita Barros, Felipe Godinho, Fernanda, Taiane Pimenta, Lorena Argolo, Nayara, Yasmin Athanasio, Camilla, Breno, Vinícius Albuquerque, Samuel Pimenta, Mateus Porto, Marcos Assunção, Rangell Guerra, Estáudio Calheiros, Késia Zaiden, Lucas Cavalcante, Gabriel, Matheus Olímpio, Thiago Matos, Débora Azevedo, Agatha Cristinny, Cinthia Cortes, Kayo Eduardo, Nylla Cristina.

2006: Renato Ferreira, Eudóxia Alice, Taiane Pimenta, Afonso Neri, Débora Azevedo, Débora Kelly, Marina Barbosa, Suellen, Samuel Pimenta, Ítalo, Ricardo Siqueira, Guilherme Araújo, Marcelo Eduardo, Douglas Nunes, Madelaine, Pedro Vaz, Tainá Fernandes, Fernanda Duarte, Anah Paula Bernardino, Roberto Arnaldo, Thaís Oliveira, Rychard, Reginaldo Sousa, Thuany Danielle, Marcos Murilo, Dayanne, Gabriel Teodoro, Gilmar, Graziela.

2007: Elias Barbosa Jr., Samuel Pimenta, Renato Ferreira, Pâmella, Larissa Melo, Mariana Nascimento, Raíssa Rocha, Jônatas Cocentino, Juliana Silva, Luana Gomes, Mayara Mariano, Marcos Moábio, Brayan Leandro, Anderson Barros, Rafael Roriz, Hugo, Banpro Nunes (*in memoriam*), Lucas Cavalcante, Natã Terra, Madson Abitbol, Yasser El-Harim, Thiago Linhares, Letícia Spezani, Úrsula Sangaleti, Fabíola, Christiane Mesquita, Wanessa Lacerda, Gemine Costa, Fernanda Rollemberg, Jéssia Moreira, Luis Alberto Rodrigues, Josiane, Gabriela Rodrigues, Carlos Trufini, Carlos Kairo Sarah Camargo, Rian Gomes, Lucas Ogliari, Arthwilliams Gomes.

2008: Christian Maciel, João Zago, Rychard Oliveira, Anah Paula Bernardino, Patrícia Leal, Melissa Trindade, Isabela Cardoso, Jéssica Dayane, Bárbara Virgínia, Felipe Pereira, Luma Mascarenhas, Jéssica Oliveira, Ana Seganfredo, Larissa Carvalho, Ângelo Lenza, Thiago Resende, Juliano Sant'Anna, Matheus Veras, Victor Leitinho, Renan Davidson, Elcimar Pereira, Suellen Maria, Thalita Barros, Daniella Sousa, Carolina Mattos, Estáudio Calheiros, Marcus Carvalho, Victor Silvestre, Alan Alves, Andressa Holanda, Fernanda Rollemberg, Gabriel Paiva, Luis Alberto Rodrigues, Victor Matheus; "Gaúcho", Ernani "Toscana"; Breno Custódio, Horicam Vítor, Taiane Abreu, Gabriela

Corassa, Matheus Cordeiro, Rafael Melo, Guilherme Lawall, Felipe Passos.

2009: Khiara Dias, Jéssica Ferreira, Gabriel Tamiozzo, Rodrigo Gonçalves, Thiago Dias, Rafael Alves, Matheus Camelier, Alisson Vinci, Yago Sávio, Moisés Paiva, Felipe Guilherme, Luiz Sampaio, Jônatas Cocentino, Fernando Fellows, Larissa Sousa, Bruna Macêdo, Carine Bastos, Dállety Kathleen, João Wesley, Talita Mariáh, Renata Visoná, Gabriella Andrade, Mateus Reis, Ariel Angel, Marina Barbosa, Thuany Danielle, Evelyne Malta, Esther Birenbaum, Igor Martins, Ananda Gonçalves, Douglas Wallyson, Leonardo Nascimento, Milena Coelho, Thaís Oliveira, Rafael Saldanha, Arthur Arrelaro, Helam Sobrinho, Esdras Aristides, Isabela Mainieri, Victor Hugo Rosa, Narrara Santos.

2010: Wanderson Barbosa, Matheus Gonçalves, Christian Maciel, João Zago, Sérgio Thadeu, Anieli Monteiro, Bruna Mendes, Bruna Lima, Ana Seganfredo, Letícia Spezani, Gabriela Ferraz, André Rabelo, Vanessa Rodrigues, Nathália Zelaya, Dáletty, Laíssa Verônica, Ana Cristina, Débora Azevedo, Marcus Carvalho, Diogo Queiroz, Kallycia Bose, Pâmella Duarte, Gemine Costa, Hiziane Ferreira, Fernando Aquino, Juliano Sant'Anna, Felipe Henrique, Arthur Monteiro, João Paulo Szerwinski, Pedro Barcelos Ariel Angel, Carlos Melo, Isabella Soares, Juliano Sant'Anna, Khiara Dias, Marina Barbosa, Patrícia Leal, Samuel Pimenta, Tainá Fernandes, Thaís Oliveira, Bárbara Costa, Marlos Chaves, Thomaz Gontigio, Roberta Madeiro, Bruna Guedes, Marcus Xavier, Daniela Carvalho, Alexandre Lenza.

I thank all friends and colleagues from the Exatas prep school, which I attended between 2008 and 2010: Marília Morais, Caio Ninaut, Jennifer Cavalcante, Artur Koberlus, Iohan Struck, Victor Camargo, Pedro Ubatan, Filipe Caldas, Ivy Caldas, Leandro Benetti, Daniella Valentim, Brenda Natália, Mariana Ramalho, Natália Huang, Gisele Spindola, Evelyn Dias, Tatiana Rodrigues, Isabela Resende, Érika Saman, Raquel C., Amanda Queiroz, Felipe Nery, Sandy Gonçalves, Saulo Kaminski, Maria Elizabeth, Raquel M., Breno, Rita, Raíza, Geovana, Kira, Mariana S., Tâmara, Anderson Dogoby, Layanne, Nudina, Isadora, Hércules, Gabriel C., Amanda M.C., Lucas Nicotti. The prep course was a big part of my routine before the great entrance exam, for a quite long time I thought contemporary colleagues as competitors or even "enemies" in a way. Life always teaches you in the best manner, and I'm grateful for my failure in 2010/2, a defeat that ultimately was a huge victory, for it has shown me that I has my true "competitor"; instead of thinking of surpassing the others, the right way was to improve myself, do my best, be humble and to always respect all others. It was a harsh lesson, but it was vital for my personal maturation. However, even in my "competition era" I made really good friends with whom I had nice experiences and a lot of fun. Although I was never a true fan of the "stand-up comedy way of teaching" commonly seen in prep school classes, I did incorporate a lot of this side in my own teaching style; even in the most serious of occasions, a nice (and smart!) joke can help as much as the best possible conceptual

explanation.

I thank all drivers and colleagues that shared rides with me during the everyday transportation to school: Hamilton, Leila, João Wesley, Loyanne, Débora, Gabriel Fernandes, Ana Luísa, Joyce, Luan Gomes, Anderson Luan, Adriana, Lucas Silva, Jonathan, Yasmim Aymé, Matheus, Nathália Zelaya, Natalia Marion, Sérgio Thadeu, Luiz, Viviane, Daniella, Carlos Eduardo, Ricardo Vieira, João Rubens, Ezgui Savaş, Igor Costa, Roberta Câmara, Aloma, Higor, Caio Augusto, Gabriel Rico, Eduarda Rollemberg, Bruno Martino, Bárbara Martino; Carlos Araújo, Vinícius Silveira, Talita Oliveira, Eglay Moreno, Gabriel, José Paulo, Bruna Póvoa, Inácio Rodrigues, Thaynara, Leonardo, Thamires, Samara, Carolina Faria, Bárbara Fernanda, Jéssica Ximenis, Jéssica Oliveira, Jade, Yasser Abdallah, Williane Ferreira; Cristiane, Raimundo, Ailton, Adilson, Álvaro Pagé, Robson, and Zezinho. Even though the musical taste ranged from bad to unbearable at most times, the comes and goes were full of random and fun stories that I keep with great affection; even the most ridiculously awful lyrics often awaken the nostalgia in me when I come to listen to them again.

For the friends and colleagues at Wizard/Guará II, in which I spent two and a half years learning French: Teachers Cláudia Lima, Ivana, Hugo, and Eduardo; João "Luanda" Baldaia, Liszt Baldaia, Jade Petersen, Yuri, Ruth, Juliano, Marcelo, Eduardo, Leonardo, Bruno, Thaynara. Most of our acquaintanceship was based on playing online games, movies, and useless kid-matter conversations, but isn't that also just a part of life... Je suis aussi reconnaissant par tous ça, vous m'avez donné beaucoup de raisons par me rappeller de ces experiences avec un sourire.

For the friends and colleagues from the Speedcubing world: Rafael Cinoto, Carlos Alcântara, Caio Lafetá, Marília Lafetá, Gabriel Bucsan, Pedro Guimarães, Rinaldo Pitzer, Alysson Dias, Ânia Gomes, Éder dos Santos, Rafael Sekia, João Pedro Batista, Wesley Dias, Diego Meneghetti, Renan Cerpe, Marcella Queiroz, Yuri Vasconcelos, Igor Butter, Pedro Roque, Fábio Bini, Israel Machado, Alaor Reis, and Arthur Arrelaro. Speedcubing was important to develop some of my motor coordination, which was always one of my main weaknesses, I especially thank Gabriel Tamiozzo for having taught me this "way to stimulate cognitive reasoning in an entertained manner", I'm kinda proud to have helped spread this "addiction" at the school to fairly "pandemic" levels in 2009... I will probably never find enough time to practice to the same "freakish levels" of high school, but I will certainly carry this as a hobby, alongside all the memories associated with this activity.

I would like to thank many people that composed a part of my life in various circumstances: Gabriel Túlio, Rafael Seixas, Artur, Renan, Alex, Felipe, Victor Squilli, Vinícius Lino, Francisco, Leo, Mateus, Humberto Conrado, Santiago Maria, Jean Guilherme, Gabriel B., Luiz O., Bruno F., Moisés C., Adriano, Andrey, Eduardo, Carlos, Max, Kaleb Kuya, Coach Marcos, Penha, Gabriel, Caio, Kim, Victor, Lucas, and Coach Capela;

Camila, Arthur, Matheus T., Wesley P., Paulo Marion, Robson, Rafael Ávila, Keoma, Ulisses, Any, Thiago, Flávia, Chelsea, Juscelino, Marilda, Raimundo, Alexandre, Simão Pedro, Vanair, Pedro Of., Lucimara, Luci, Rosângela, Antônio T. (*in memoriam*), Édio, Rosilene, João Ricardo, Fábio, Lúcio, Carla, Cláudio, Lúcia, Rodolfo, Jéssica, Pietro, Alessandra, Michelle, Larissa, Helliane, Naiara, Pedro, Arthur, Felipe, Fernanda, Marcela , Alan Régis, Édipo Consoli , Rosângela, Gal, Elmo, Conceição, Francisca, Matias, Laurindo Teixeira, Cida, Gleuton, "Bicho", Kaynan, Maurício, Paula, Gabriel A., Gabriel C., Brenda, and Kim; Suami, Felipe, Evandro, Evandro Jr., Jane, Edna, Luiz Ambrósio, Ferreira, Vivinane, Victor Hugo, Max George, Tiago Correia, Roger, William, Rebecca, Juan, John, Eric, Luiz Carlos, Adriano, Alessandra, Manuel Cruz e Sousa, Alessandra, Marcus, Samantha, Yuri, Marilúcia, Byron (*in memoriam*), Derby, João Pedro, Glória, Carolina, Diego, Thiago, Vladimir, Miriam, Valmir, Arivaldo, Mariângela, Gisela, Julita K., Juçara, Vera, José Vaz, Edson, Nilza, Tiago M., Heloísa, and Lenita; Jorge Massabane (*in memoriam*), Edna, Anderson, Norma, Esther, Milton, Luiz Carlos, Marcos P., Marcos J., Marcos K., Magda, Aristos, Vanessa, Paulo, Robertson, Silvio, Elaine, Caio, Nair, Benedito, Timóteo, Manuela, Augusto Flumm, Marcos Nezão, Teresa R., Ricardo, Luciana, J., J. Jr., Paulo Vianna, Herenice, Otaviano, Elizabeth, Tereza, Marisa, Jônatas, Minami, Evaldo, Berenice, Marcus F., Ângela, Gilmar, Sueli, Lincoln, Priscilla, Isângelo, Isandson, Marcos R., Cleide, Michelle, Andressa, Rubens G., Stella (*in memoriam*), Wellington M., Uébio, Beto Burgos, Marta, Jonas, Romário, Inocêncio (*in memoriam*), Márcia, Gildo, Mônica, Faustino, Cristiane, Thiago, Silas, Ari, Vera, Thierry, Vitória. A man should never forget his origins, I had memorable moments with all of you, some joyful and entertaining, some bitter and painful; nonetheless, regardless of the motivations or the implications, those moments defined me, and for that I surely recall all of them with great satisfaction.

As a scientist, I realized that science alone doesn't make you a better person, nor build a better world. The greatest scientific discoveries can bring forth the worst calamities when combined with bad intentions. All systems converge to an equilibrium, nature's wisdom is far greater than men's. What goes around comes around: cliche as it might sound, understanding this is quite a shortcut to conquer a lot of things.

What good are technicality and precision if one achieves them after losing the essence that defines his existence... Sometimes small things are the key to make a better world. Desires tend to infinity, so they are unachievable by definition; perhaps "happiness" is to have the courage to say "That's enough" and learn to be grateful to all apparent imperfections.

Be happy, and don't lose yourself. That basically wraps it up for me. I can't think of anything better I could give to someone, so I hope this little "abstract" can make your future better somehow.

Abstract

This thesis discusses the emergence of Big Data and machine learning and their applications in various aspects of Business Administration, emphasizing the methodological contributions of this inductive-based paradigm in finance and the improvements of this approach over econometric tools and traditionally well established methods of data analysis. The statistical foundations of machine learning are introduced and the challenges of high-dimensionality in finance problems are analyzed, including the practical implications of nonlinearity incorporation, regularization of the additional complexity level and forecasting for high-frequency data. Finally, three empirical applications are proposed, concerning respectively on volatility forecasting, portfolio allocation, and stock price direction prediction; in those applications, different machine learning models are explored, and the insights from the results were discussed in light of both the finance theory and the empirical evidences.

Keywords: Machine learning; Complexity; Nonlinearity; Bias-Variance Dilemma; Finance.

Resumo

A presente tese discute a emergência do *Big Data* e do aprendizado de máquinas em vários aspectos da administração de empresas, enfatizando as contribuições metodológicas deste paradigma baseado no raciocínio indutivo em finanças e os benefícios desta abordagem em relação a ferramentas econométricas e métodos tradicionais de análise de dados. Os fundamentos estatísticos do aprendizado de máquina são introduzidos e os desafios da alta dimensionalidade em problemas financeiros são analisados, incluindo as implicações práticas da incorporação de não-linearidades, a regularização do nível de complexidade adicional e a previsão em dados de alta frequência. Finalmente, três aplicações empíricas foram propostas, relativas, respectivamente, à previsão de volatilidade, à alocação de portfólio e à previsão da direção do preço de ações; Nessas aplicações, diferentes modelos de aprendizado de máquina foram explorados, e os *insights* dos resultados foram discutidos à luz da teoria financeira e das evidências empíricas.

Palavras-chave: Aprendizado de máquinas; Complexidade; Não-linearidade; Dilema viés-variância; Finanças.

List of Tables

List of abbreviations and acronyms

ANN	Artificial Neural Network.
CAPM	Capital Asset Pricing Model.
GARCH	Generalized Auto-regressive Conditional Heteroskedasticity.
LASSO	Least Absolute Shrinkage and Selection Operator.
MAE	Mean Absolute Error.
OECD	Organisation for Economic Co-operation and Development.
PCA	Principal Component Analysis.
RMSE	Root Mean Square Error.
RMT	Random Matrix Theory.
SFFS	Sequential Forward Floating Selection.
SVM	Support Vector Machine.
SVR	Support Vector Regression.
TS	Tournament Screening.

Contents

1 Introduction: Big Data and machine learning in business administration

Mankind has always been searching for answers to explain the unknown and to understand the purpose of his existence: throughout the history, people have been listing unsolved questions and analyzing evidences collected from empirical observations, organizing and processing those "data" with both logic and intuition and ultimately using them to make assumptions and formalize theories trying to explain the world and the society, as well as the role of the individuals that compose them. Therefore, the concept of taking data to propose solutions is not at all new. However, the development of technologies and the "speeding-up" of the everyday human routine led to an abundance of data without precedents, which culminated in the so-called "Big Data era".

Data has become "Big" in many different ways: data are increasingly "taller", as the number of observations stored in databases grows as more and more people get connected to online services and social media; analogously, data are increasingly "fatter", as the number of features associated to each individual also grows due to the aforementioned integration of sources of information; finally, data are increasingly "faster", as the spreading of breaking news and oscillations of financial indicators can already trigger significant social and economic implications faster than a person can even react to those updates. Concerning the pillars that constitute what is Big Data, White (2012) defined five dimensions: besides the volume of the data, the information come in a big variety of forms and sources, are transmitted in a high velocity, have the potential to generate business value, and come in uncertain degrees of quality and veracity.

The very existence of more abundant information in all those "dimensions" allows scientific researchers from various knowledge fields to tackle existing problems from different perspectives, to propose new models based on variables that were previously not available, to develop new analytical tools and methods of data analysis to cope with the larger complexity in the data, as well as to pose new questions based on new insights derived from the analysis of this "bigger data". As pointed out by the work of Sivarajah et al. (2017), there is an increasing interest of both scholars and market practitioners for using Big Data Analytics as a tool to aid strategical decision-making, by boosting the operational efficiency and assuring competitive advantages in the market. Furthermore, the paper performed a systematic literature review mapping relevant publications about the implementation of different types of Big Data solutions for organizations, discussing their contributions – both academic and empirical – to the use of these novel technologies in managing organizational resources.

While the availability of massive data opens up new possibilities for scientific inquiries, the real challenge is to extract the relevant information out of the "pile of data" and convert them to useful outputs to be used to knowledge construction, to propose solutions to relevant problems, and to aid real-world decision-making. For instance, the paper of Gandomi and Haider (2015) discussed the dominance of unstructured data over structured ones, as the vast majority of data come in videos, audios, and texts that are not "tidy", thus demanding not only an initial work of cleaning the data but also a wide set of statistical tools to extract non-spurious relationships and efficient computational algorithms to analyze them in a feasible time.

Similarly, based on case studies and a systematic literature review, Wamba et al. (2015) proposed a general categorization of the role of Big Data in generating business value, summarized into five broad dimensions: 1) Transparency creation; 2) Experimentation enabling to improve performance; 3) Population segmenting; 4) Full or partial automation of human decision-making; and 5) Innovation of business models and services. Moreover, the authors listed issues related to the business value enabled by the presence and development of Big Data, indicating that the access to data and the introduction of new techniques and technologies are the most debated issues in recent scientific productions.

As discussed in Boyd and Crawford (2012)'s work, "With Big Data, comes Big Responsability". Alongside the abundance of data, the ethical implications of using them also come into account, as well as the implications of the widespread integration between different sources of information in the formation of a potential new digital divide defined between the availability of big data and the scarcity of the agents who actually have access to them. In this sense, Big Data presents itself as more than a series of technical advancements, thus being a broader social phenomenon instead. Foster et al. (2016)'s work also discuss the implications of data errors in the conclusion drawn by analyzing them, emphasizing thus the importance of recognizing the limits of Big Data tools when applied to real world problems, specially in social sciences. In this sense, the artificial "intelligence" cannot fully replace the human intelligence, serving as a complement of it instead.

The intersection of statistics, computation, and social sciences – often jointly known as "data science" – has been consistently growing, and this trend is also seen in various spheres of the organizational environment (CHEN; CHIANG; STOREY, 2012). As pointed out by the paper of Sagiroglu and Sinanc (2013), the application of data science to corporative problems can lead to advantages over competing firms but also demands more sophisticated solutions for the storage and the analysis of the data, as well as challenges about information privacy and business ethics. The work of Brynjolfsson and McAfee (2011) also addresses the integration between data science and businesses,

and in which sense nowadays organizations have a unprecedented opportunity to accurately measure the profiles of their customers and make experiments using those data with machine learning algorithms to make innovations over existing technologies and improve their services and relationship with their clients.

About the emergence novel technologies and the escalating complexity of the interactions between the data, Wang et al. (2018)'s paper briefly describes the evolution of technology and the gradual paradigm shift of its role in the society, conceptualizing the idea of a "parallel society" – also called Society 5.0 – as a subsequent step of the "network society" and a transition to a "Cyber-Physical Social System", in which the social connections would be managed by knowledge automation.

Concerning the automation of tasks and the potential replacement of human workers by machines or automatons, the book of Brynjolfsson and McAfee (2014) argue that the technologies that arose in the past decades is already enough to actually execute many tasks as good as a human would do, and pointed the implications of this trend to the economy and social welfare. As estimated by Frey and Osborne (2017)'s research, in the next years the technologies that will become available can make nearly half of the jobs in the United States to be automated; other works like Frey et al. (2016) and Bosch, Pagés and Ripani (2018) reached similar conclusions, reporting estimated percentages of automation as high as 85% for developing countries with less diversified economies like Ethiopia and Guatemala. Indeed, the results are consistent with the findings of Arntz, Gregory and Zierahn (2016), which analyzed the automation risk for 21 OECD countries and reported that workers with larger educational levels and jobs that demand higher technical qualifications were less prone to automation, although many occupations and the labor market itself would likely to undergo deep transformations with the presence of robots and automatons performing not only repetitive and non-cognitive tasks, but also showing the potential to "learn" new knowledge and provide analytical insights based on the experience gained from the received data.

Regarding the limits of "artificial intelligence" in scientific discoveries and the role of human researchers in a "algorithmic era", Titiunik (2015)'s work states that the availability of a larger volume of data alone is not able to induce a paradigm shift in causal inference inquiries; instead, the previously established theoretical background is essential to make sense into the results "mined" from the Big Data, as well as adequately-planned research designs to draw valid conclusions out of the empirical evidences. As affirmed by the author: "There are no algorithmic or automatic shortcuts to scientific discovery. In fact, the need for critical thinking will be stronger the more we become inundated with ever-larger amounts of new information".

In this sense, the rise of "Big Data" and the prominence of machine learning methods have the potential to reshape the *modus operandi* of the scientific research itself. As

pointed out in the paper of Gelman et al. (2011), while the deductive approach focuses on testing a previously established theoretical framework, the inductive approach focuses on updating those theories after observing potentially new patterns from the data. Given the flexibility of machine learning techniques regarding assumptions like the functional form of the models and data distribution, they fall into the second category and aim to extract patterns from the massive volume of "Big" data available that might have gone unnoticed under a more restrictive model structure. The paper of George, Haas and Pentland (2014) discuss the implications of using the patterns extracted from the "big data" to explore causal relationships, reaffirming the validity of using the inductive paradigm to complement existing theoretical constructs – built with under a deductive approach – through the empirical experiments. Moreover, the authors emphasized that the presence of many sources of unstructured data opens a very wide spectrum of plausible interpretations and subsequent causal explanations between the variables, arguing that this plurality reinforces the importance of both previously established theories and evidences collected from new observations that test the theoretical assumptions in the empirical realm.

Therefore, the inductive approach can identify patterns that can ultimately be used to improve gaps that the deductive theories fail to explain, consequently allowing to build better theories that will then be retested with new data, and so on. The influences of the boosting of computational power in statistical findings and causal inferences are discussed in Efron and Hastie (2016)'s research, which stated that the use of inductive tools – such as machine learning algorithms – in scientific inquiry represent not only a paradigm shift towards a broader way to solve the proposed research question but also an analytical framework that allows breaking assumptions that do not necessarily hold in real-world data, and potentially "discover" recurrent patterns and stylized facts whose causal explanations go beyond the scope of existing theories.

Concerning the use of Big Data to generate value for firms and organizations, Günther et al. (2017)'s job argue that while breakthrough advancements have taken place in the last few years in data mining and business intelligence solutions, the organizations have not yet adapted themselves to those transformations in various levels of analysis not limited to the debate of the interactions between human and artificial intelligences, but also regarding organizational business models, access control of the data and the trade-off between value generation and negative social externalities. The authors then proposed an integrated model for value realization with big data influenced by the portability (possibility of transferring and applying data) and the interconnectivity (possibility of integrating many data sources into a common structured framework). Indeed, machine learning algorithms can be applied to a wide scope within business administration, not only in the sense of solving problems efficiently and providing support for decision-making, but also as a mechanism to deal with challenges that arise with Big Data by adapting

to them, transforming the organizational environment and actively building competitive advantage, as summarized in the sections below:

1.1 Big Data and machine learning in public administration

As stated by the work of Lavertu (2016), public administration is rapidly transforming as a consequence of the speeding-up and scaling-up of data collection and analysis, which the author calls a "Big Data revolution" that brought alongside it both opportunities of improvement and risks for the management and evaluation of public programs and policies, as the intensified dissemination of high granularity data allowed external actors – often with little project-specific expertise – to influence the policy evaluation process, which can, in turn, deviate the priority of active government actions into costly and non-efficient decisions. Analyzing recent data from primary and secondary education from the United States, the study concluded that the improvement of governance is essential to cope with the rapidly increasing use of Big Data in public organizations.

Chen and Hsieh (2014)'s paper also pointed out that Big Data represents one of the most prominent challenges for governments in the digital era, analyzing its potential to promote efficient utilization of information and communication technologies, as well as improving online public services, ideally towards personalization according to the profile of the receiving citizen. Besides, the study made a case study of a Big Data implementation initiative in Taiwan summarizing the key challenges in data management, emphasizing the need of a solid, stakeholder-focused and performance-oriented governance structure, as well as the assurance of digital privacy and security. Practical applications include Marzagão (2015), which developed an app that uses natural language processing to classify products purchased by the Brazilian government into classes taking the products' description as input data, intending to reduce misclassification and subsequent bad spending of the public budget.

A common challenge in public policy evaluation is to estimate the so-called "treatment effect" of a government intervention in those who actually received it, and compare this effect to its counterfactual – what would the outcomes be if those same observations had not received the treatment, such as a social assistance or a healthcare program. Regarding this kind of problems, the report of Athey (2015) pointed out that methods of supervised machine learning can be useful to analyze not only the relationship of the outcomes with features held *cæteris paribus* when the intervention takes place, but also the variation of the causal effects on different settings of those features. The same author, in Athey et al. (2019), also proposed a similar method for estimating the counterfactual values of outcomes for the treatment group for a panel data setting.

A synthesis of the value of Big Data and machine learning for governments – and

how this value can be propagated for businesses and scholars – can be found in the work Kim, Trimi and Chung (2014), which discussed the possibility of converting both structures databases and unstructured data into projects and solutions that may lead to better wealth distribution, enhance the operational efficiency and the budget transparency, or even boost the engagement of citizens on the country's economic performance and national security. A similar debate can be found in Desouza and Jacob (2017)'s paper, which drew insights about decision-making in public administration based on interviews performed to public sector Chief Information Officers from federal, state and local levels of the United States regarding the planning, the execution, and the implementation of public sector Big Data projects. A broad discussion of the use of Big Data in the public sector, as well as its potential benefits and costs, is present in Maciejewski (2017)'s work, which presented cases and feasible applications in policy design and in public internal management.

1.2 Big Data and machine learning in marketing

Erevelles, Fukawa and Swayne (2016)'s paper analyzed the transformation of marketing with Big Data, from its collection and storage to its conversion into patterns and consumer insights, and finally being actively used to boost organizational capabilities and generate competitive impacts. Furthermore, basing the analysis on resource-based theory, the authors argued that in the digital era, firms tend to benefit more from inductive reasoning than from deductive reasoning, such that Big Data and radical innovation like the adoption of machine learning can be used to create value and competitive advantage. Finally, the paper concludes that one of the core factors to truly make the most of Big Data is the creative intensity, in the sense of converting the available technologies to solutions that meet the demands of this new era. Ethical implications of the usage of Big Data – such as privacy incursions in recommender systems and invasive marketing – are discussed in works like Boyd and Crawford (2012).

The research of Glass and Callahan (2014) discussed the implications of Big Data and digital marketing in corporate strategies and organizational competitiveness, analyzing case studies and listing future trends in marketing management considering the multiple sources through which a potential customer may interact with an organization brand with the emergence of Big Data, new technologies and overlapping social networks. Indeed, machine learning models can provide valuable inputs for research agendas such as neuromarketing: the use of neuroimaging in marketing and business can be seen in works like Ariely and Berns (2010), while machine learning methods like convolutional neural networks have already achieved solid advancements in image recognition and pattern extraction; in this sense, a joint analysis of a customer's visual activity and its purchase profile can aid corporate managers in issues such as product placement, publicity planning,

and personalized marketing campaigns.

The abundance of geospatial data and the prospects of using them in management are debated in Karimi (2014)'s article, exploring technical issues in geospatial data collection and innovative solutions such as geo-crowdsourcing, as well as applications in business analytics and social media management. A branch named geomarketing combine the geospatial dimension in to the models, allowing to build decision support systems that may help a manager to geographically allocate franchises in a way that maximizes the potential economic payoff taking into account features such as overall income level and consumption patterns, aiming a more directed impact over a target public that is compatible to the goals of the organization. Recent applications that combined machine learning methods with geospatial analysis in management sciences include Oliveira (2016) and Padula et al. (2017).

1.3 Big Data and machine learning in logistics and supply chain management

The paper by Waller and Fawcett (2013) examined the bridges between supply chain management with data science, predictive analytics, and Big Data, reinforcing the increasing popularity of this intersection and listing a set of desirable quantitative skills and theoretical background to not only tackle challenging research questions, but also being able to convert them into intelligible courses of action for managers and implementable tools to improve management performance. The study also listed potential applications of Big Data in logistics regarding issues like inventory management, determination of optimal routes, and sentiment analysis; moreover, the authors exemplified research questions relevant to supply chain management grounded on many classic management theories, such as contingency theory, agency theory, and institutional theory.

As stated in Wang et al. (2016a)'s article, machine learning can provide deeper insights concerning market trends and consumption patterns, being applicable as well to efficiently manage maintenance cycles and minimize costs in operations management. The authors indicated that Big Data Analytics can be not only descriptive or predictive, but also prescriptive in the sense of being able to impact different capability levels of supply chain management, such as the process efficiency and processing speed. A compendium of techniques to collect, disseminate and analyze Big Data in logistics and supply chain management was listed and defined as a set of strategic assets that can be integrated across different business activities.

Gunasekaran et al. (2017)'s paper investigated the impact of the assimilation of Big Data and predictive analytics on supply chain and organizational performances based on a resource-based view: "assimilation" of said Big Data tools were defined as a poste-

rior step of acceptance and routinization, influenced by the resources – sharing and the connectivity of tha information. The results showed a positive cycle between Big Data Analytics acceptance and the resources, mediated by the commitment of the top management, an effect that spreads positively to supply chain and organizational performances. An application can be seen in the work of Zhong et al. (2015), in which authors proposed a data warehouse of data enabled by radio frequency identification, coupled with a dimensionality reduction framework, to determine spatio-temporal logistic paths whose efficiencies are quantitatively measurable; the authors then discussed the potential usage of this "holistic Big Data approach" in the planning and scheduling of a firm's logistic operations, as well as many managerial implications of its implementation.

1.4 Big Data and machine learning in human resources management

Bearing in mind the relevance of extracting knowledge from Human Resource data, Ranjan, Goyal and Ahson (2008)'s work address the importance of data mining methods and techniques in Human Resource Management Systems, and in which extent those methods can determine the company's competitive position and organizational culture. The authors presented many potential applications of machine learning in finding measurable patterns for human resources management, including: classification of best résumés in recruiting processes, identification of employees with top performance or higher probability of leaving, identification of groups of attrition, predicting the behavior or attitude of the workers, finding the best designation of tasks to a group of employees, amongst many others. Therefore, the paper shows that data from human resources are still under-explored, whilst the application of data mining algorithms have the potential to not only improve the quality of the decision-making process, but also convert the data into increased performance, satisfaction at work, and competitive advantage for the organization.

In light of the development of technologies and the potential of jobs automation, the research of Davenport (2014) analyzed the transformations on the relationship between the workers and those technologies, as well as a broader discussion concerning the nature of the jobs and the role of organizations in this new era. A similar discussion can be found in Veloso et al. (2018)'s manuscript, which analyzed the perceptions of Brazilian management students about pressure from new technologies on their career perspectives, as well as the adoption of new technologies (such as social networks) on learning and working environments. The paper concluded that the incorporation of new technologies induce a heterogeneous perception depending on the nature of the work and the tasks involved, with professionals on operational positions feeling more pressured than ones in

management positions, emphasizing thus the lesser risks of automation for tasks that are intensive in cognition and latent factors such as charisma, sensibility, and leadership.

Hecklau et al. (2016)'s work also investigates the effects of technologies of Industry 4.0 in redefining competences demanded by organizations and the need for new strategic approaches for human resource management. Specifically for manufacturing companies, the automation trend for simpler and more repetitive tasks projects an increase of more complex workspaces, thus demanding higher levels of technical qualification of the employees; in this sense, the paper presents the challenge of tuning the workers' qualifications to make them able to execute more sophisticated processes and to ensure the retention of jobs in a rapidly changing labor market. Recent studies that analyzed the shifting of jobs towards competencies that are less likely to be automated include Bosch, Pagés and Ripani (2018) and Albuquerque et al. (2019).

1.5 Big Data and machine learning in business innovation

Viewing the Big Data era as a interlacement of connections between machines in a growing collaborative community, the study of Lee, Kao and Yang (2014) discussed the need of innovation and transformation in traditional manufacturing services, especially concerning the development of scalable computational tools to manage the massive information that flow in and to convert the data into assets that help to manage uncertainties and provide more robust decisions. The authors classified automation-centered manufacturing system and service innovations are "inevitable trends and challenges for manufacturing industries", thus being key elements for productivity and transparency for organizations immersed in this new reality. Gobble (2013)'s paper made a similar diagnosis, emphasizing the rapid acceleration of data volume over an already frenetic pace.

The work of Huda et al. (2018) addressed the swift evolution of information and communication technology in recent years and the use of Big Data and digital devices like tablets and smartphones in educational applications, for both teachers and students, encouraging a design thinking for innovative ways of virtual learning; moreover, the authors indicated a promising research agenda concerning the possibility of integrate Big Data concepts and methods into online learning, allowing the personalizing of pedagogical strategies for each student based on their Internet behavior and preferences, which can be achieved using machine learning algorithms like clustering, natural language processing and sentiment analysis to potentially enhance the students' performance and overall development.

As discussed in Carayannis, Sindakis and Walter (2015)'s research, the configuration of the business models is a key feature that define corporate strategies, as well as their

understanding and implementation, both within an organization and between them in the market. In this sense, business model innovation can induce effects such as governance efficacy organizational sustainability. In this sense, the investigations reported by Sorescu (2017) pointed out the prominence of business model innovation over product innovation in the Big Data era, as seen from the practical experience of many successful startups over the recent years, addressing the ways in which organizations can leverage highly connected information networks – regarding both internal and external data sources – to innovate over their business models; in special, a business model innovation does not necessarily have to be radical or disruptive, even it actually manages to generate a high amount of value.

Similarly, the study Schüritz and Satzger (2016) proposed a framework of "data-driven business model" focused on Big Data and Analytics and the ways an organization can transform itself to get inserted into this new paradigm. The paper states that the integration of Big Data into the organizational decision-making process opens up a wide range of transformation prospects for business models. instead of restraining into mere tools to solve specific problems. The study identified five different patterns in which Big data can modify the business model, and based on a sample of 115 cases of companies that publicly stated the usage of Big Data Analytics solutions, the paper concluded that a "New Data-Infused business model" provides an innovative architecture with a potential driving force for not only value creation, but also value capturing and value proposition. Whilst this new paradigm is still scarce in corporate environments, the study reinforces the potential of Big Data and the necessity of going beyond of simply collecting the data, but also being able to identify opportunities of using them and actually converting the models and algorithms into feasible revenue boosts for the company.

1.6 Big Data and machine learning in finance

Forecasting is a relevant issue in finance. Along many years of empirical verification, numerous scientific works have identified some patterns that occur consistently with financial data, patterns commonly known as "stylized facts", as summarized in works like (CONT, 2001). Those stylized facts show that financial time series exhibit, amongst other behaviors, non-stationarity of the prices over time, non-constant conditional variance, clustering volatility (*i.e.:* Periods of high volatility tend to also be followed by high volatility, and the same for low volatility periods), data not following a Gaussian distribution, etc. Therefore, the search for better forecasting tools for financial data has remained a prominent research agenda, yielding a high number of scientific productions proposing many different approaches for obtaining the best prediction based on a given sample data. Machine learning techniques fall in as one of those approaches: its use has been consistently increasing over the recent years, motivated not only by the good

out-of-sample forecasting performance but also representing an emergent paradigm with potential to bring new critical reflections over many well-established results in financial theory.

For instance, one of the most well-known postulates in finance is the efficient markets hypothesis, introduced by Fama (1970)'s paper, which states basically that the price levels of financial assets tend to the equilibrium in which there is no optimal trading strategy that can systematically beat the market; the market eventually converges to the equilibrium price level that may incorporate the information of past data ("weak-form"), all publicly available data ("semi-strong-form") or even the inside information ("strong-form"). Nonetheless, studies like Gerlein et al. (2016) and Ramakrishnan et al. (2017) present empirical evidence that the use of machine learning techniques in financial trading decisions can yield out-of-sample predictions capable of "beating the market" for a wide set of market segments and training periods.

In special, one key feature of this class of models is their ability to introduce a high degree of nonlinearities into the explanatory variables. For instance, a widely used algorithm called "Support Vector Machines" (introduced by Cortes and Vapnik (1995)'s research) can cope with nonlinear interactions by means of a Kernel function, which can actually map the original data into an infinite-dimensional feature space with a small number of parameters.[1] The advantages of nonlinear relationships for financial forecasting is discussed in the work of Hsu et al. (2016) and will be further and accordingly explored in posterior sections.

The article Varian (2014) categorized data analysis in econometrics into four main steps: prediction, summarization, estimation, and hypothesis testing – and argued that while traditional econometric tools focus on the economic significance of the statistical estimates, aiming to identify potential causality relationships, machine learning methods provide better ways to predict and to generalize. In this sense, the development of the computational power can aid economists and social scientists in general to extract knowledge from non-intuitive relationships that may not appear when considering models structures that are easy to interpret; in fact, the paper presented a simple example showing that even simple machine learning models like the decision trees are able to easily introduce nonlinearities that models like logistic regression fail to capture.

Nevertheless, a major setback in introducing nonlinearities is to keep them under control, as they tend to significantly boost the model's complexity, both in terms of theoretical implications and computational power needed to actually perform the calculations. Nonlinear interactions, besides often being difficult to interpret, may bring alongside them, apart from a potential better explanatory power, a big amount of noisy information – *i.e.*, a increase in complexity that is not compensated by better forecasts or

[1] For technical details, see Schölkopf and Smola (2002)'s book.

theoretical insights, but instead "pollutes" the model by filling it with potentially useless data.

Bearing in mind this setback, the presence of regularization is essential to cope with the complexity levels that come along with high dimensionality and nonlinear interactions, especially in financial applications, in which the data-generating processes tend to be highly chaotic. While it is important to introduce new sources of potentially useful information by boosting the model's complexity, being able to filter those information, discard the noises and maintain only the "good" information is a big and relevant challenge. Works like the one from Massara, Matteo and Aste (2016) discuss the importance of scalability and information filtering in light of the advent of the "Big Data Era", in which the boost of data availability and abundance leads to the need to efficiently use those data and filter out the redundant ones.

Finally, this thesis addresses the implications of a high dimensionality not only "by columns", in form of nonlinear decision functions, by "by rows" as well, which would be translated into the use of high frequency data in financial applications. While this approach is not particularly new, combining machine learning and regularization methods with the analysis of the behavior of financial time series in smaller time frequencies can reveal patterns that would not emerge in low frequency data, allowing a further understanding of the complexity, the degree of "chaos" and potential predictability of financial phenomena.

In light of the relevance of the discussed topics, this thesis analyzed the relationships between the use of machine learning methods in finance, the importance of introducing new orders of complexity into the predictive models and being able to keep them under control. Specifically, applications and the potential advantages of the use of nonlinear interaction between the predictors and high frequency data in finance studies are presented; the mathematical foundations behind the bias-variance dilemma and insights from chaos theory and econophysics are also discussed. A broad discussion concerning the statistical implications of complexity and a literature review of recent works who dealt with high dimensionality in financial applications are displayed in chapter **2**. Moreover, three applications are proposed in chapters **3**, **4** and **5**, concerning respectively: 1) volatility forecasting for traditional currencies and cryptocurrencies and their behavior in low and high frequencies; 2) the effects of noise filtering in portfolio selection; and 3) the impacts of feature selection in stock price prediction. For each application, the theoretical background were discussed based on the specialized literature and the methods applied on the empirical analysis were separately structured.

2 Complexity, regularization and machine learning: Challenges of high dimensionality in finance

2.1 Nonlinearities and machine learning in financial applications

Machine learning models are characterized to be "inductive" in a sense that they are flexible to the data collected in the sample, yielding decision functions based on the patterns that the data show, instead of fixing functional forms (like a linear or quadratic one) or assuming presuppositions about the distribution of the data. So, instead of assuming that the financial data are normally distributed or homoscedastic, which are premises empirically shown to be not true, and force the data to a fixed framework knowing that the results may be distorted due to the incompatibility of the assumptions, machine learning techniques does not demand such kind of assumptions besides assuming that the sample taken is representative – a postulate which, if violated, would provide invalid estimates even in traditional econometrics. For a compendium of the use of machine learning models for financial data and evidences their empirical desirability, see Sewell (2017).

In a recent paper, Nakano, Takahashi and Takahashi (2018) built trading strategies for bitcoin – a notoriously volatile asset – using predictions of its price level based on a deep artificial neural network approach and tested their profitability for different sets of transaction cost, measured by the magnitude of the bid-ask spread. The authors used a seven-layer neural network with rectified linear unit (ReLU) as activation function and managed to surpass the buy-and-hold strategy in terms of profitability between December 2017 and February 2018, a period in which the prices bitcoin suffered severe drawbacks. These results show that, under certain circumstances, the increment in predicting power brought by machine learning techniques can be strong enough to not only outperform the market but generate economic profits by exploring the oscillations of the market, thus displaying positive evidences toward the application of this class of methods.

One of the key features of machine learning methods, and pointed by many scientific works as one of the main sources of forecasting power boost over traditional econometric models (KANAS, 2005; CONRAD; LAMLA, 2010; CHAO; SHEN; ZHAO, 2011; BURNS; MOOSA, 2015), is the introduction of nonlinearity into the data-generating process and into the explanatory variables themselves. More often than not, the relationships between the specified variables and the effects they induce occur in a nonlinear way, which

makes the use of linear models in these contexts generate potentially biased results that can decisively hinder the decision-making process (MAMA, 2017). As discussed in the paper of Han et al. (2017), even though linear functional forms bring alongside many mathematical conveniences and the practicality of being easily interpretable, escalating the data into a high-dimension level can allow the researcher to better understand and explore underlying functional relationships across different dimensions of interactions that would simply be ignored analyzing the data only linearly.

Similarly, Brock (2018a)'s work broach the subject of introducing high-dimensionality into financial data analysis summarizing methods of financial time series analysis that are able to detect patterns that reflect into deterministic chaos behavior in the low dimensional realm, including machine learning applications. The study concludes that the identification of nonlinear interactions can substantially improve short-run forecasting for time series and that the majority of the conventional statistical methods fail to consider those interactions, thus arguing in favor of the use of the former approach over the latter ones.

The study of Brock (2018b) makes the link between the fields of economics and finance with concepts of the theory of complex dynamics, viewing the financial market as a broader complex system than the deterministic chaos patterns that emerge in low dimensions. The author postulate that the attractor set to which the dynamics of the financial markets converges is not a single point, but rather a spectrum of hidden states with no limit steady state. Furthermore, the evidences of the complex dynamics still hold even when eliminating sources of stochastic exogenous shocks. Bringing this idea in view of the financial theory, it implies the existence of a "dynamic equilibrium" in which the "equilibrium" traditionally defined in finance theory (for instance, in CAPM and Efficient Markets Hypothesis) is one of the many "equilibria" to which the complex system converges and the one that emerges when considering low dimensionality data and linear functional forms. Conversely, when introducing high dimensionality into the explanatory variables and the data-generating process' functional form – for example, by using nonlinear predictors – the additional information may point towards underlying "equilibria points" in the short-run, which can help to understand and ultimately leads to the equilibrium defined in the "linear world".

Buonocore et al. (2016)'s paper present two key elements that define the complexity of financial time series: the multi-scaling property – which refers to the dynamics of the series over the time; and the structure of cross-dependence between time series – which are reflexes of the interactions among the various financial assets and economic agents. In a financial context, one can view those two complexity elements as systematic risk and idiosyncratic risk, respectively, precisely the two sources of risk that drives the whole motivation for the risk diversification via portfolio allocation, as discussed by the Modern

Portfolio Theory.

It is well known that the systematic risk cannot be diversified. So, in terms of risk management and portfolio selection, the main issue is to pick assets with minimal idiosyncratic risk, which in turn, naturally, demands a good estimation for the cross-interaction between the assets available in the market, namely the covariance between them.

The non-stationarity of financial time series is a stylized fact well known by scholars and market practitioners, and this property has relevant implications in forecasting and identifying patterns in financial analysis. Specifically concerning portfolio selection, the non-stationary behavior of stock prices can induce major drawbacks when using the standard linear Pearson correlation estimator when calculating the covariances matrix. Livan, Inoue and Scalas (2012) provides empirical evidences of the limitations of the traditional linear approach established in Markowitz (1952), pointing out that the linear estimator fails to accurately capture the market's dynamics over time, an issue that is not efficiently solved by simply using a longer historical series.

The heterogeneity of financial time series correlation structures patterns induced by the stylized facts have motivated many papers that analyze its implication in financial markets; the different correlation structures over time and their evolution over time were analyzed in many financial applications, such as mapping the states of a financial market (MÜNNIX et al., 2012).

In light of evidences that not all noisy information of the covariance matrix is due to their non-stationarity behavior (MARTINS, 2007), methods like Support Vector Machines (GUPTA; MEHLAWAT; MITTAL, 2012), Gaussian processes (PARK et al., 2016) and deep learning (HEATON; POLSON; WITTE, 2017) have been discussed in the literature, showing that the introduction of nonlinearities can provide a better display of the complex cross-interactions between the variables and generate better predictions and strategies for the financial markets. For an overview of the applications of machine learning techniques in portfolio management contexts, see Pareek and Thakkar (2015).

Musmeci, Aste and Matteo (2016) incorporate a metric of persistence in the correlation structure between financial assets, and argue that such persistence can be useful to anticipate market volatility variations and quickly adapt to them. Testing for daily prices of US and UK stocks between 1997 and 2013, the correlation structure persistence model yielding better forecasts than predictors based exclusively on past volatility. Moreover, the paper discusses the effect of the "curse of dimensionality" that arises in financial data when a large number of assets is considered, an issue that traditional econometric methods often fail to deal with. In this regard, Hsu et al. (2016) argues in favor of the use of non-parametric approaches and machine learning methods in traditional financial economics problems, given their better empirical predictive power, as well as providing a broader

view well-established research topics in the finance agenda beyond classic econometrics.

2.2 The drawbacks of complexity: Hoeffding's inequality and regularization

Thus, to allow a certain degree of in-sample error is not necessarily a bad thing, since the main goal is to make generalizations based on a representative sample taken from the population. To effectively make the in-sample error to be zero is not a difficult task; however, in doing so, one would be implicitly assuming that the patterns observed in that particular sample would occur for future out-of-sample data not observed yet. In this way, the algorithm would be simply memorizing the past data – not only the actually relevant information but also the noisy information particular to that specific sample – which in turn can hinder its predictive power.

The work of Wolpert and Macready (1997) presents a result concerning search and optimization, suggestively called "no free lunch theorem" by the data science community, which states that, in statistical inference, the aggregate cost-benefit relationship of all candidate solution methods (*i.e.:* The joint factor of each method's computational cost and solution performance) is equal for all problems to be potentially solved. This implies the non-existence of a globally superior learning algorithm that dominates over the others for every application – in other words, there is no "best" model that suits well every application; instead, there is the right method for the right application. In this sense, not all "nonlinearities" are equal, as shown in many empirical applications: for example, in Henrique et al. (2016)'s paper on portfolio allocation using Support Vector Regression, the inverse multiquadric Kernel function yielded the best results, while in Yaohao and Albuquerque (2019) this function yielded overall poor results for exchange rate prediction.

Thus, since different "clans" of nonlinearities produce different results, it is natural to think that the introduction of nonlinearities, *per se*, does not necessarily mean that the predictive performance will go up. While considering nonlinear interactions allow to capture a whole new class of patterns that may be informative to the empirical inference, the introduction of noisy and uninformative data may actually jeopardize the generalizing ability of the proposed models. Rising up the dimensionality of the model may be useful, but doing so unmanneredly can fill that high dimensionality with noises that compromise the model's quality and usefulness. Therefore, in the same way that the introduction of nonlinearities is a relevant issue to be tackled, being able to keep the additional complexity of the model under control is equally fundamental.

But how can one find the optimal middle ground between a model that fits well enough to the in-sample data with high components of nonlinearity and a model not excessively noisy? In other words, to what extent the additional complexity brought alongside

a high dimensionality model does more good than harm and how can the researcher actively control it? A result called *Hoeffding's inequality* (HOEFFDING, 1963) provide a possible answer to those questions:

Hoeffding's inequality is basically an exponential version of the more well known Chebyshev's inequality, which provides an upper bound for the probability of a random variable X with arbitrary distribution to have a certain distance to its mean. [1]. Similarly, Hoeffding's inequality provides the probability for the distance between the in-sample error E_{in} and the out-of-sample error E_{out} to be less or equal than a user-specified margin $\varepsilon > 0$: its expression is given as:

$$\mathbb{P}(|E_{in}(g) - E_{out}(g)| > \varepsilon) \leq 2 \cdot \mathbb{M} \cdot e^{-2 \cdot \varepsilon^2 \cdot n} \tag{2.1}$$

where n is the sample size and $\mathbb{M}$ is a measure of complexity of the model, measurable by the Vapnik-Chervonenkis dimension of the model (VAPNIK; LEVIN; CUN, 1994), which basically expresses the learning capability of a statistical learning algorithm – *i.e.:* the space of functions that a learning method can produce as a decision function for the input data. Evidently, the broader this set of "attainable functions" to be potentially learned, the bigger will be the model's capacity to yield good predictions for out-of-sample data. However, a big space of functions also means additional complexity to be considered; therefore, in a scenario in which a "simpler" function would suffice, considering more complex functions to be the potential optimal solution would be a weakness, since it would be simply adding more noise and unnecessary complexity, thus hindering the generalization ability of the model.

Besides providing a probabilistic upper bound for the generalization error of a model's decision function, Hoeffding's inequality formalizes the trade-off between capacity and complexity for the construction of a good algorithm for generalizations. A good model for this purpose is one that lies on the optimal middle ground between describing well the data taken from the sample, as well as deriving patterns for future and yet unseen data that are not way too complex, since the past data is filled not only with useful information but also an intrinsic component of noise, such that merely "memorizing" the past data tend to be not enough to cover satisfactorily future predictions. In statistics, this trade-off is also known as the bias-variance dilemma.

Consider $\mathcal{H}$ a set of "candidate functions" from which one function will be selected by the learning algorithm as the best predictor for future data. The comprehensiveness of $\mathcal{H}$ will determine the model's overall complexity (one of the possible measures is the Vapnik-Chervonekis dimension). Assuming that in-sample data is composed by n observations of p variables $\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_n}$, the in-sample error associated to each function

[1] Algebraically, Chebyshev's inequality states that for any random variable X with finite expected value μ and variance σ^2, it is valid that $\mathbb{P}(|X - \mu| \geq k \cdot \sigma) \leq \frac{1}{k^2}$, for any real number $k > 0$.

$h(.) \in \mathcal{H}$ can be easily obtained, while the out-of-sample error depends on future data $\mathcal{X}$. Assuming a classification problem, those components can be defined as:

$$E_{in}(h(\boldsymbol{x_i})) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(h(\boldsymbol{x_i}) \neq f(\boldsymbol{x_i})) \tag{2.2}$$

$$E_{out}(h(\boldsymbol{x_i})) = \mathbb{P}(h(\mathcal{X}) \neq f(\mathcal{X})) \tag{2.3}$$

where $\mathbf{1}(.)$ is the indicator function. For a regression problem, the in-sample error may be defined in many ways, such as the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE):

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|h(\boldsymbol{x_i}) - f(\boldsymbol{x_i})| \tag{2.4}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(h(\boldsymbol{x_i}) - f(\boldsymbol{x_i}))^2} \tag{2.5}$$

While E_{in} can be effectively reduced to zero without major challenges, doing so tend to be harmful to generalization purposes, since the researcher would be basically "hoping" for the future data to be equal to the sample taken. On the other hand, the generalization error E_{out} depends on the allocation between in-sample bias and out-of-sample variance (complexity). After algebraical manipulations of the expression **2.1**, the upper bound for E_{out} can be expressed as:

$$\begin{aligned} E_{out}(h(\boldsymbol{x_i})) &\leq E_{in}(h(\boldsymbol{x_i})) + \sqrt{\frac{1}{2\cdot n} log\left(\frac{2\cdot\mathbb{M}}{\delta}\right)} \\ &\leq E_{in}(h(\boldsymbol{x_i})) + \Omega \end{aligned} \tag{2.6}$$

where $\delta = 2\cdot\mathbb{M}\cdot e^{-2\cdot\varepsilon^2\cdot n}$ is a constant and Ω represents the penalization for the complexity of the model. Note that there are two sources that raise the upper bound for generalization error E_{out}: A model does not generalize well if the sample data are not well described or if the model is way too complex to fit well for unseen data. From a purely mathematical point of view, the problem of a decision function's excessive complexity can be seen in the behavior of the deviation between the Lagrange polynomials interpolations (WARING, 1779) and the actual function it is interpolating: the approximation error grows larger for higher degrees of said polynomial – a problem named "Runge's phenomenon" (RUNGE, 1901).

Machine learning methods, while being flexible to the characteristics of the data, bring alongside a negative feature called **overfitting**. As the name implies, overfitting

occurs when the predicting algorithm ends up in describing too well the in-sample data incorporating not only the relevant data-generating process but the noises specific to that particular sample as well.

A simple example, illustrated by figure 1, can evidence the problems with the two extrema that jeopardize generalizations. Linear regression, although being a model with very low complexity (asymptotic behavior of the decision function in very simple), clearly does not fit well to the data, resulting in a model with low E_{out}, but high E_{in}; on the other hand, the degree 15 polynomial regression shown in the sub-figure on the right fit very well into the in-sample data, but it follows along the noises of that sample, and ends up not approximating well the true population function intended to predict.

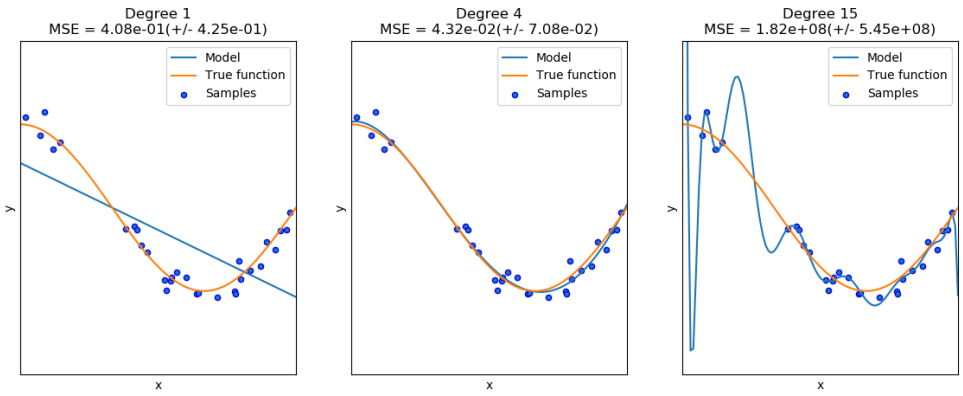

Figure 1 – Underfitting vs overfitting (SCIKIT-LEARN, 2017)

Since machine learning methods are capable to cope with high degrees of complexity – especially in terms of introducing nonlinear interactions between the explanatory variables – methods that deal with the excessive complexity are needed for the models to not over-fit. Such techniques are known as regularization methods. In machine learning, typically the data is split between training and test sets so that the algorithms are forced to be subjected to new data to check whether they indeed "learned" the relevant patterns using data from the training set. However, there are many other regularization techniques designed to control the model's variance. Specifically in finance, many papers use regularized models and often achieve equal or even better forecasting performance with a less complex model. As states in "Occam's razor" principle, a simpler explanation with same explanatory power tend to be the best one.

Gu, Kelly and Xiu (2018) applied various machine learning methods – namely principal components regression (PCR), partial least squares (PLS), generalized linear models (GLM), boosted regression trees, random forests and artificial neural networks –

and compared them to simple and penalized (ridge regression, LASSO and elastic net) linear models to measure the risk premium of financial assets using data of nearly 30000 financial assets from New York Stock Exchange and NASDAQ. Using monthly data from 1957 to 2016, the authors fitted regularized models modifying their objective functions – instead of optimizing only the loss function, many versions of penalty terms for complexity were added. The results presented empirical evidences favoring machine learning models in terms of providing a more accurate description of the price oscillation patterns of the analyzed assets in comparison to traditional statistical methods; the authors credit the predictive performance gain to the introduction of nonlinear predictor interactions from those methods, which are not considered by commonly used econometric approaches. Moreover, the authors reported that all models converged to a similar set of "essential predictors" composed mainly by variations in the assets' liquidity and volatility, arguing in favor of parsimonious models over complex ones. The use of regularization managed to successfully shrink the model's variance without curtailing the predictive performance in the same proportion of the penalizations. In particular, in this paper's application, the "deep" architecture neural network showed worse results than "shallower" networks, possibly due to the high levels of noise present in financial data, and also showing that, in certain circumstances, a simpler and more regularized model may actually perform better. This topic will be further explored later alongside the discussion on the potential contributions of Random Matrix Theory to finance.

Feng, Polson and Xu (2018) propose a nonlinear feature extraction to map the most informative components that explain the patterns of financial assets over time, treating the sorting of securities as an activation function of a deep neural network. Using return data of US stock market asset between 1975 and 2017, the authors show that the well-known Fama-French models with three (FAMA; FRENCH, 1992) and five factors (FAMA; FRENCH, 2015) are actually particular cases of the proposed deep learning approach; those factors were compared the "deep factors", with the deep learning cases slightly outperforming the 3-factor case, but showing higher mean square error than the 5-factor case, while ordinary least square showed high levels of out-of-sample error, again evidencing the importance of controlling the ideal level of complexity added to the predictive model.

Barfuss et al. (2016) emphasize the need for parsimonious models by using information filtering networks, building sparse-structure models that showed similar predictive performances but much smaller computational processing time in comparison to a state-of-art sparse graphical model baseline. Similarly, Torun, Akansu and Avellaneda (2011) discuss the eigenfiltering of measurement noise for hedged portfolios, showing that empirically estimated financial correlation matrices contain high levels of intrinsic noise and propose several methods for filtering it in risk engineering applications.

In a financial context, Ban, Karoui and Lim (2016) discuss the effects of performance-based regularization in portfolio optimization for mean-variance and mean-conditional Value-at-Risk problems, showing evidences of its superiority towards traditional optimization and regularization methods in terms of diminishing the estimation error and shrinking the model's overall complexity.

Concerning the effects of high dimensionality in finance, Kozak, Nagel and Santosh (2017) tested many well established asset pricing factor models (including CAPM and Fama-French 5-factor model) introducing nonlinear interactions between 50 anomaly characteristics and 80 financial ratios up to the third power (*i.e.:* all cross-interactions between the features of first, second and third degrees were included as predictors, totaling models with 1375 and 3400 candidate factors, respectively). In order to shrink the complexity of the model's high dimensionality, the authors applied dimensionality reduction and regularization techniques considering ℓ_1 and ℓ_2 penalties to increase the model's sparsity. The results showed that a very small number of principal components are able to capture almost all of the out-of-sample explanatory power, resulting in a much more parsimonious and easy to interpret model; moreover, the introduction of additional regularized principal components does not hinder the model's sparsity, but does not improve predictive performance either.

2.3 High frequency financial data and forecasting

In financial time series analysis, usually the standard periodicity of the data is the daily frequency for common applications such as stock market prediction and volatility estimation. Nonetheless, in recent years the literature has been moving towards increasingly higher time frequencies or even a volume based paradigm so that the daily data is used as a low frequency baseline to which the higher frequency data are compared. Easley, Prado and O'Hara (2012), for instance, discuss the emergence of a "volume-orientated paradigm" for financial analysis, which focuses on developing forecasts and trading strategies considering a high frequency data horizon. Given the sharp increase in worldwide financial transaction flows, scholars and financial market agents tend to progressively shift to the "volume clock" – based on the number and volume of financial transactions –, instead of the traditional "time clock", which follows the chronological flow and sets the minimum variation interval to a fixed amount of time. Hence, the traditional baseline daily frequency is increasingly less capable of coping with the number of transactions in a single time period, emphasizing the relevance of a high frequency trading paradigm for nowadays finance.

In addition to being strongly influenced by recent events or the availability of market information, as discussed by Reboredo, Matías and Garcia-Rubio (2012), high

frequency data boost the dimensionality that a researcher will be dealing with. For example, while a sample with daily data for a quarter of a year yield only 90 observations, a small number considering its splitting into in-sample and out-sample subsets, as the usual procedure for fitting machine learning algorithms. However, by taking hourly data, the sample for the same time periods would have 2160 observations, a number large enough to evoke many asymptotic theory results, such as convergence to a Gaussian distribution.

Furthermore, analyzing the high frequency data can reveal nuances that may pass by unnoticed by taking a smaller time frequency. For example, the impacts of news may cause fluctuations to the price level of a financial asset but be fully incorporated within a day, such that the daily variation (and consequently, the log-return) would not capture such oscillation. Using a candlestick chart, one could notice the relative volatility level induced by the exogenous shock to that specific day, albeit still prone to miss potentially relevant details. Using intraday data, however, would allow the researcher to observe the details of the price adjustment dynamics, which may lead to the identification of underlying patterns arising from the additional information incorporated by the model. For a discussion about the informational gain of high frequency trading and the construction of trading strategies using those kinds of data, see Gomber and Haferkorn (2015).

Due to the importance of fresh news, regarding assets price, (ANDERSEN; BOLLERSLEV, 1998) explained that financial series present an extremely volatility behavior since they incorporate expectations and reactions of economic agents in the face of events. Currently, market asset volatility forecast and estimation are highly relevant in the composition of derivatives prices, in the portfolio risk analysis and in the investment risk analysis itself. So, the development of methods that help decision making arouses great interest among investors.

Camargo, Queiros and Anteneodo (2013) made a segmentation analysis of the trading volume and the price oscillations with minute frequency data for the 30 companies indexed at the Dow Jones Industrial Average during the second semester of 2004. The study identified a slow decay in the autocorrelation function between these two factors, suggesting a persistence over time that a linear estimator is not capable of incorporate. Moreover, the authors indicate that the volatility of price tends to be higher than the trading volume for a given time span, which implies that the market's adjustment dynamics tend to be faster than investors can assimilate them. On the other hand, the trading frequency in financial transactions has been increasing over the past years, which motivates a high frequency paradigm for financial forecasting and trading strategies, as described by Easley, Prado and O'Hara (2012).

Aloud et al. (2013), using tick-by-tick market data of the Euro/US Dollar currency pair between 2007 and 2009, also present empirical evidences of high correlation between the intraday price volatility and the trading volume; moreover, the paper concludes that

the trade volume and numbers of closing/opening positions are shown to be scale-invariant to the price variation threshold. This finding motivates the analysis on the effects that large price variations bring upon the covariance structure of the financial assets, and whether the traditional linear estimator suffers from it. Therefore, we also verify the inconsistency of linear covariance estimation in high temporal frequencies, as well as the effects of introducing nonlinear interactions in the portfolio's overall profitability.

Back to the expression of Hoeffding's inequality, it was mentioned that the generalization error can be defined as the aggregation of the in-sample error and the model's complexity. It basically states that the best model is one that makes the most of the information present in the training sample, without exploiting it with an excessively complex decision function. Thus, since the incorporation of high frequency data can provide ulterior information over low frequency data, it is expected that that information come alongside a component of noisy and non-informative features, which may pollute the intended analysis if the predictive algorithms were to over-fit for the training sample, being specifically harmful to forecastings in the traditional low frequency data horizon. For example, a researcher can be interested in estimate financial volatility for daily frequency and decide to use intraday data for a source of additional information, following the idea of Heterogeneous Autoregressive models (HAR), as introduced by Corsi (2009). The additional information provided by intraday can theoretically improve the accuracy of the forecasts, and indeed it does empirically, as shown in works like Wang et al. (2016b) and Tian, Yang and Chen (2017). However, since the HAR model uses aggregate intraday volatility as an independent variable to estimate the daily realized volatility, articles like Audrino and Knaus (2016) point out that the gains of HAR models over well-established models like GARCH and its extensions arise from the fact that HAR's structure shrinks off part of the complexity of financial information carried on over time, especially in view of the stylized fact of their non-stationary behavior. In this way, the aggregation of time frequencies contributes positively to the model's predictive power due to the specification of a more parsimonious model – a smaller degree of complexity to be penalized, viewing from the Hoeffding's inequality – using intraday information which, theoretically according to the efficient market hypothesis, would be fully incorporated in the daily data. This represents a gain in a way of diminishing the model's overall complexity while being able to approximate long memory dependence, hence lowering the out-of-sample generalization error.

Works like Audrino and Knaus (2016) pushes this idea further, by performing LASSO regularization in the HAR model to further eliminate its complexity. The study concludes that the predictive performance of HAR with LASSO is statistically equal to its non-regularized counterpart, whilst being more parsimonious and converging to the traditional lag structure of HAR if the latter is indeed the true model. Audrino, Huang and Okhrin (2016) shows similar results, discussing the limitations of the traditional HAR

specification, which fixes a lag structure of (1,5,22) for the daily volatility (corresponding to lags for the previous day, five working days in a week and twenty-two working days in a month). The conclusions show that peaks of instability into the financial market makes the HAR lag structure not adequate, while the adaptive LASSO framework allows the model to "learn" the best specification to avoid biased estimates or overfitting.

Moreover, the regression framework called Mixed-Data Sampling (MIDAS), developed originally by Ghysels, Santa-Clara and Valkanov (2004), also explores this idea: the basic specification for this kind of models involve a predictor measured in a higher time frequency than the dependent variable. In this way, the MIDAS approach works as an alternative formulation for state space models, which works on the idea of updating the forecasts based not only on past information but also on the arrival of new observed outcomes. Bai, Ghysels and Wright (2013) showed in their work that the MIDAS regression specification is a general case for the Kalman filter (KALMAN, 1960) when considering mixed frequency data between independent and target variables; the main difference is that Kalman Filter is estimated through a system of equations, while MIDAS reduces the structure down to a single equation. Golosnoy, Gribisch and Liesenfeld (2012) developed a Conditional Autoregressive Wishart model to estimate the covariance matrix for five assets listed in the New York Stock Exchange, reporting that the model managed to capture long-run oscillations the time-dependence structure by incorporating both MIDAS and HAR methods into the estimators.

Marsilli (2014) discusses the potential setbacks of combining different time frequencies using the MIDAS model: while the aggregation of high frequency data can allow one to extract more information out of the sampled data, the choosing of the predicting variables remains obscure and have a decisive influence on the resulting estimates. This author applied dimensionality reduction techniques into MIDAS and proposed two variable selection criteria, namely by a LASSO-penalizing MIDAS and by Bayesian stochastic search. The results showed that the methods managed to successfully identify key informative predictors for forecasting US economic growth rate mixing low (monthly) and high frequency (daily) data. The idea of using high frequency data to model long-run persistence appears also in Bollerslev, Patton and Quaedvlieg (2016), which proposes to incorporate the variation of the magnitude of the volatility measurement errors over time, adding a bandwidth of variation for the estimated parameters.

3 The best of two worlds: Forecasting high frequency volatility for cryptocurrencies and traditional currencies with Support Vector Regression

Abstract

This paper provides an evaluation of the predictive performance of the volatility of three cryptocurrencies and three currencies with recognized stores of value using daily and hourly frequency data. We combined the traditional GARCH model with the Machine Learning approach to volatility estimation, estimating the mean and volatility equations using Support Vector Regression (SVR) and comparing to GARCH family models. Furthermore, the models' predictive ability was evaluated using Diebold-Mariano test and Hansen's Model Confidence Set. The analysis was reiterated for both low and high frequency data. Results showed that SVR-GARCH models managed to outperform GARCH, EGARCH and GJR-GARCH models with Normal, Student's t and Skewed Student's t distributions. For all variables and both time frequencies, the SVR-GARCH model exhibited statistical significance towards its superiority over GARCH and its extensions.[1]

3.1 Introduction

Some controversy among financial economists lean on the highest precision on volatility estimation. Merton (1980) and Nelson (1992) noticed that the volatility forecasting doesn't need a huge amount of historical data, instead, a short period of observation is enough to make such analysis (ANDERSEN et al., 1999).

It was also observed that, with an arbitrarily short span of data, it is possible to get an accurate volatility estimation (POON; GRANGER, 2003). For this reason, the progress of volatility studies are related to the use of higher frequency data. Regarding that, this work provides an evaluation of the predictive performance of the volatility of three cryptocurrencies and three traditional currency pairs, using daily and hourly frequency data.

In order to estimate volatility, researchers most use the GARCH model. However, nowadays the Support Vector Regression (SVR) emerged as a strong and robust

[1] Published in *Expert Systems with Applications*, v. 97, p. 177–192, 2018

method, capable of covering multivariate and dynamic characteristic of financial series. This method is rooted in the Structural Risk Minimization (SRM) process, which aims to estimate the nonlinear data generating process trough a risk minimization and a regularization term to achieve the minimal unknown populational risk.

In this context, we proposed the application of this study using the cryptocurrency market. These new assets have a new combination of characteristics that makes them so unique in operation and transaction, been unable to relate completely with others markets for several reasons: first, compared to the commodities market, they do not have a great historical background and no future market to be the benchmark, but despite that, we were able to achieve an interesting result in the volatility forecast. Second, the use of cryptocash is different from the traditional cash; such that even if a country adopt it as an official digital currency, the transactions were designed to be done directly between economic agents without the need of an intermediary institution, monetary control or accountability system. Third, the cryptocurrencies value and distribution is based on a P2P-network, it has no physical representation to handle it, only a string is necessary, which is called a wallet, and its password, that is used to send and receive cryptocurrencies. For those reasons it is important to do a specific research in the cryptocurrencies market, using traditional and novel methodologies to create a model capable of understand the unique characteristics and dynamics provided by this new asset.

This research presents a combination of SVR approach with a GARCH model (SVR-GARCH) and tested it against several traditional GARCH models and well known extensions. Furthermore, the machine learning based model was tested for both cryptocurrencies and traditional currencies, in order to check whether this approach yields significant boosts in predicting ability, as well as investigating potential similarities and differences between cryptocash and traditional money. Moreover, we replicated the tests for low and high data periodicities and for different time periods, in order to verify whether SVR-GARCH model's predictive performance is satisfactory over the whole time extension of the series, using Diebold and Mariano (1995) test for predictive accuracy and Hansen, Lunde and Nason (2011) Model Confidence Set.

This paper's contributions consists in joining two emerging research agendas in finance; the study of cryptocurrencies and the use of machine learning forecasting techniques: while numerous papers presented applications of machine learning based models in various research agendas in finance, works that link this approach to cryptocurrencies remain scarce. Specifically, cryptocurrencies' volatility levels are usually much higher than traditional currencies (YERMACK, 2013), making prediction associated to this segment a potentially even more challenging task than to commonly addressed variables, such as stocks indexes or exchange rates. Therefore, we compared a machine learning based model to well established models in financial econometrics for both traditional and

cryptocurrencies, and verify the relative performance of the machine learning paradigm in both "worlds".

This paper is structured as follows: Section 2 presents key features of cryptocurrencies, discussing their relevance in finance and comparing them to traditional currencies. Section 3 describes the theoretical background of volatility estimating methods in high frequency data, as well as the benchmark models and the methodology used to estimate volatility in this paper. Section 4 addresses the empirical analysis of the daily and hourly volatility estimation using bitcoin, ethereum and dash prices (in US dollars) and the spot exchange rate between US Dollar and Euro, British Pound and Japanese Yen between January 4th 2016 and July 31th 2017. Section 5 presents the forecasting results and discuss their implication in view of the financial theory. Finally, Section 6 presents conclusions and remarks, showing limitations to this approach and recommendations to future researches.

3.2 Theoretical background

3.2.1 Cryptocurrencies

Ferguson (2008) presents a brief chronology of the evolution of the concept of "wealth": in the medieval era, wealth was associated to having the means to conquer and pillage, so that it was regarded as a consequence of power; with the rise of mercantilism and capitalism afterwards, wealth began to be interpreted as the possession of material goods (such as precious metals) and the means to generate production and trade it for more material goods; thus, money was increasingly viewed as the cause of power. As the capitalist system consolidated in the western society, the main indicator of wealth became the possession of money, since its high liquidity allows it to be converted into any other asset. Nowadays, the hard core of the world's wealth is concentrated in financial assets rather than real assets: at this stage, a successful businessman's fortune is mainly evaluated based on his company's stock value, instead of his yacht or his luxurious car.

While a consolidate model is still in discussion, in 2009 a new kind of asset appeared in the market, the bitcoin, leading the world to analyze this newcomer and try to figure out its place and dynamics, and whether the rise of cryptocurrencies can change the concept of "wealth" once more. Recent works like Vigna and Casey (2016) argue that the new ideas and technologies that come alongside with cryptocurrencies, such as the blockchain and the decentralization of money, have the potential to lead the world into a "new economy": while a "revolution" on the foundations of social life is unlikely to occur, the introduction of decentralized cybermoney tend to push the traditional ways of economic transactions even further into the digital realm, nerfing the cost of global scale transactions, which in turn provides viability for individuals to work for companies in

different countries, and makes even well-established companies to adapt their corporate strategies into this new reality.

Bitcoin wasn't the first digital currency: e-Gold (G&SR, 1998), eCash (CHAUM, 1983), Beenz (COHEN, 1998) and Flooz (LEVITAN, 1999) were previous attempts of a purely virtual way to make transactions, with e-gold being the most successful among them. Created by Gold & Silver Reserve Inc. in 1996, e-Gold was an anonymous service that allowed the possibility to make instant transfers of value – ownership of gold and other precious metals – to other e-gold accounts (G&SR, 2006a), transaction flows reached a peak of U$2 billion per year envolving over a million users (G&SR, 2006b). E-Gold attracted great attention, but with the possibility of user-anonymity, cybercriminals, money-launderers and other kinds of criminals developed interests for that service and a series of prosecutions occurred against it. In 2008, the CEO of Gold & Silver was sentenced and all e-Gold's accounts were frozen, the company had to close a few months later and the end of e-Gold was declared. Other digital currencies had a similar end, by similar causes, but an alternative would be born sooner.

Satoshi Nakamoto (NAKAMOTO, 2008) proposed bitcoin in order to be an easy way to make transactions over the Internet, which works globally, faster, independent from an institution to operate it and limited to 21 million units, such that only the deflation is expected to occur (DARLINGTON, 2014), a feature that may be used as a vanish point for citizens "running away" from economies that suffer great levels of inflation (VASQUEZ, 2017). Bitcoin represented a breakthrough over the previously cited digital currencies for solving cryptographic problems like the "Trusted Third Party" (TTP) (ANDRYCHOWICZ et al., 2014) and the "Double-spending" (ROSENFELD, 2014). Thus, bitcoin would solve the failure points of older attempts, giving the market a much more solid solution to the market aspirations for digital currencies.

Bitcoin is based on a peer-to-peer ledger that is governed by mathematical restrictions, this ledger only allows real transactions to be written in it. Every new transactions go to a pool of "unconfirmed transactions", while miners take these transactions and write them into the longest chain of verified blocks. Every new block linked to the block of a mined transaction is called a "confirmation"; the suggested number of confirmations is 6 (COMMUNITY, 2017). If a criminal intends to fake 6 confirmations, it will cost him more then half a million dollars in bitcoins reward for mining the block (as of November 2017), apart from the huge computation power required to this. So, the faking process would not yield economic gains for transactions lower than this "faking cost". Moreover, the attempt of writing fake transactions must outdo the computational power of half the miners to write the longest blockchain, which is impracticable both economically and computationally. In fact, since its inception, bitcoin never registered a single falsification with more than 1 confirmation. Every transaction and the agents involved are regis-

tered, any change over the bitcoin data nullify its uses in future transactions. Until now the average number of transactions registered are 288,155 per day (COINDESK, 2017a; BLOCKCHAIN, 2017a), a growth of 40% compared to the same period of 2016, in which registered around 205,246 transactions per day.

Many key concepts of bitcoin are relatively new to financial theory: bitcoin has no association with any authority, has no physical representation, is infinitely divisible and is built using the most sophisticated mathematics and computation techniques. For those reasons, cryptocurrencies can give fear to any specialist that is not used to this kind of "money" and is widely astonishing to the market world. Different from the traditional currencies that operate in the market, the bitcoins and other cryptocurrencies that follow similar logic don't have their value based on any country economy or in some physical and tangible asset, like the gold-dollar parity under the Bretton Woods system. Instead, the conception of value has it origin in the security of the algorithm, traceability of the transactions and the precedence of each bitcoin. Also, the exact number of circulating bitcoins in the market is known, so this asset's money supply is constant, providing a incentive for bitcoin owners to simply keep them still without trading them for goods or services, which in fact happen at a high proportion; unlike in traditional centralized currencies, in which the boost on money supply would affect the relative value of banknotes (KRISTOUFEK, 2015).

Due its nature, bitcoins assume a dual feature as a medium of exchange and as an asset of investment. Authors like Polasik et al. (2015) found both characteristics in bitcoins evident in different time windows and various market investors' profiles; others such as Evans (2014) and Segendorf (2014) discuss about bitcoin's appliance, although there are still some critics about how this asset will deal with legal matters and taxes, as pointed by Polasik et al. (2015). Evans (2014) present the concern with the bitcoin's volatility and in what extent it would affect the payments transactions made with it, and states that the market would stabilize with time or upon the formation of big wallets that would assume the role as value guarantee. According to Segendorf (2014), the analysis of bitcoin's volatility is based on the market reality and the risk that bitcoin present if used as an official digital currency, the security of information and transactions still discussed as an apprehension topic in contrast with the Swedish case as an example of functionality, making bitcoin's implications in the financial market a issue of interest for investors and scholars alike.

The debate about bitcoin's taxonomy has by itself aroused great academic interest. While the exponential increase of bitcoin's prices resembles a bubble behavior, it might not be purely related to speculative aspects, as indicated by recent academic studies: Gandal and Halaburda (2014) state that the inclusion of bitcoin into a diversified portfolio significantly increases its risk-adjusted returns, due to both bitcoin's high average returns and

low correlation with other financial assets. Bouri et al. (2016) present evidences that indicate that bitcoin can be indeed used as a hedge to market specific risk. Dyhrberg (2016a) analyzes the volatility of bitcoin in comparison to US Dollar and Gold - traditionally regarded as "safe" value reserves - using GARCH (1,1) and EGARCH (1,1). The paper concluded that bitcoin bears significant similarities to both assets, specially concerning hedging capabilities and volatility reaction to news, suggesting that bitcoin can be a useful tool for portfolio management, risk analysis and market sentiment analysis. As evidence of the recent acknowledgment of the hedge propriety of bitcoin, economic agents already invested a total of 19 billion dollars in the cryptocurrency until March of 2017, suggesting an increase of the usage, popularization and trust in the bitcoin (COINDESK, 2017b; BLOCKCHAIN, 2017b).Dyhrberg (2016a) also points out that the bitcoin reactions may be quicker than gold and Dollar, thus substantiating the analysis of both high and low data frequencies in this paper. The author replicates the study using TGARCH(1,1) and find similar conclusions (DYHRBERG, 2016b).

The other financial line of bitcoin studies is the speculation factor that comes with the high volatility, making the arbitrage possible to investors. Yermack (2013) for example criticizes the bitcoin as a currency and a hedge asset, pointing the obstacles to make it a functional digital currency, since its value and liquidity don't behave as other real currencies do, pointing out the bitcoin's speculative potential due its limited amount available in the market. Another supporter is the speculative characteristic is Kristoufek (2015): while the author recognizes the similarities between bitcoin and traditional currencies, bitcoin is shown to have a more dynamic and unstable value over time that drives its nature to be a speculative asset more than a currency.

When it comes to comprehending the market and the economic agents involved, bitcoin uses are still very restricted to some countries and activities. Usually, the one's that are open to this new asset are technological or innovation centers, which are able to understand the potential and advantages of the cryptomoney. Estonia, United Sates, Denmark, Sweden, South Korea, Netherlands, Finland, Canada, United Kingdom and Australia are ten countries friendly to the uses of bitcoin. The main application of cryptocurrencies in those countries was on the on-line marketing, known as e-commerce, in exchange for products and services. However, due the increase of bitcoin's market value, its exchange nature has been relatively put aside.

The asset flexibility in transactions (since there are few regularizations of this new market) and the high level of the cryptography gives the coin enough trust to be used instead of cash, like in Denmark. Even with the importance and its uses, the worldwide acceptance is still far from happening and the impact over the cryptocurrency dynamics in the present is inevitable. Yet, there is not enough literature review that discusses or presents efficient methodologies to estimate the new bitcoin market behavior around the

world. Some authors already conducted studies about this new market, (YERMACK, 2013; XIONG; BAO; HU, 2014; DYHRBERG, 2016a; BOURI; AZZI; DYHRBERG, 2016), but they are still restricted to the unknown aspects of the behavior and/or acceptance of bitcoin.

Bitcoin is not the only cryptocurrency, as other types of digital currency were created, called altcoins. Therefore, in addition to bitcoin, we chose another two relevant cryptocurrencies, ethereum and dashcoin, using market cap and price as criteria, to proceed with the same volatility analysis with the SVR models

Ethereum is a branch from the original bitcoin project, so it inherits his principal concepts, but ethereum was not designed to be a rival of bitcoin, it actually used bitcoin principles to produce more technology by including "smart contracts", creating a new world of possibilities, like implementing a voting system without any third party to trust, with his virtual machine an a programmable system this coins has not only promise for a revolution in the cash system but a revolution in any contractual system. Ethereum runs with no chance of censorship, fraud, third party interference or downtime, since it functions precisely as programmed (KIM, 2016) .

Since Ethereum's "smart contracts" are built with a part of an enforcing mechanism, they are considered more flexible. The major difference of Ethereum from other cryptocurrency is that it automatically enforce the clauses of an agreement if one of the participants disrespect it. It is important to mention that all the counter-party risks cannot be mitigated with these contracts and that it is complex to enforce it in certain cases (BALTA et al.,) .

Another important cryptocurrency is Dash, which comes from Digital Cash. It deals with instant transactions and is characterized as a privacy-centric digital currency cryptography, ensuring total anonymity. Also, by using a two tier network, Dash improves bitcoin system. In addition, Dash apply an anonymity technology that precludes the acknowledgment of who made the block chain, which is a similar for bitcoins. This technology uses a protocol mix utilizing an innovative decentralized network of servers, this advance in the system is known as Masternodes. This provide a more trustworthy system. (KIM, 2016).

3.2.2 Cryptocurrencies and traditional currencies

The classic definition of "money" requires an asset to be usable as a medium of exchange, a unit of account and a store of value. Researches like Urquhart (2016) indicate that cryptocurrencies such as bitcoin still present informational inefficiency, even though it has been showing a trend towards efficiency. Thus, separating the volatility analysis of bitcoin prices considering different time horizons may provide a better understanding

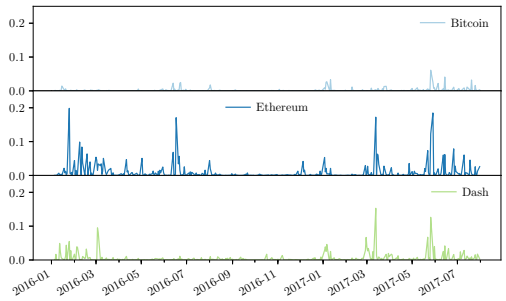

Figure 2 – Cryptocurrencies daily volatility

Figure 3 – Exchange rates daily volatility

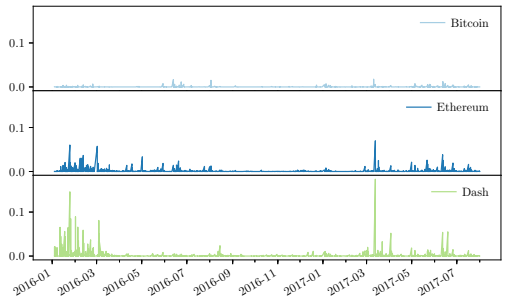

Figure 4 – Cryptocurrencies hourly volatility

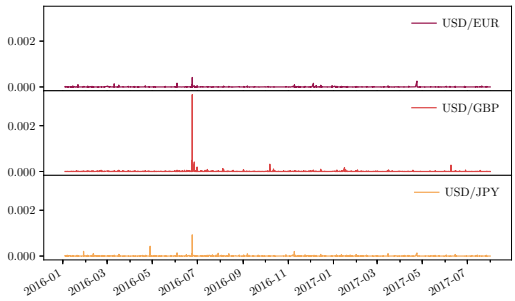

Figure 5 – Exchange rates hourly volatility

regarding this finding. Additionally, studies like Yermack (2013) and Dowd (2014) states that cryptocurrencies are susceptible to speculative bubbles that can mine its fundamental value, and their behavior characterizes them as a speculative investment rather than a currency.

On the other hand, according to authors like Lo and Wang (2014), cryptocurrencies like bitcoin has satisfied all three functions of money, emphasizing the similarities between virtual and real coins. Based on this argument and the discussion regarding the proportion of their fundamentalist value in relation to the speculative component, it is relevant to analyze the dynamics of their volatility over time. Dyhrberg (2016a) also argues that bitcoin can be classified as an asset between a "pure medium of exchange", like US Dollar, and a "pure store of value", like gold. While bitcoin is not a currency *per se*, it combines advantages of both Gold and Dollar and has the potential of being an important instrument in financial markets and portfolio management.

The reason for choosing cryptocurrencies is centered in the innovative and potential that these assets have in the current global economy. The method of machine learning itself has already presented good results in diverse situations, cases and problems, since it is possible to find a pattern and estimate decision guide lines, but for cryptocurrencies

there is the differential of unknown and unimaginable future. Cryptocurrencies bring along themselves not only potentially profitable assets, but also new concepts that differ greatly from the traditional monetary economy, like direct transaction, intangible value base, absence of a central institution to assure its store of value, as well as propelling advances in security and information storage at global scale.

As seen in the worldwide spillover effects of financial crisis in late 1990s (BELKE; SETZER, 2004), high levels of volatility can bring over a herd behavior and financial contagion that can lead to unpredictable and large turnovers in the financial market. Thus, analyzing whether the volatility patterns of cryptocurrencies behave similarly with traditional currencies can be a important framework to better understand cryptocurrencies' role in nowadays finance and investigate whether there are evidences that the "virtual world" currencies are ready to merge into the "real world".

Furthermore, issues concerning the heterogeneity among different cryptocurrencies such as the potential influence on the operating system of different cryptocurrencies on their market behavior or volatility level remains unexplored in the finance literature. Indeed, we observed that the volatility patterns of the three major cryptocurrencies we picked (bitcoin, Ethereum and Dash) differ significantly among themselves in both daily and hourly time frequencies, as seen in figures **2** to **5**. The overall volatility levels of cryptocurrencies is much higher than the exchange rates, in both daily and hourly frequencies; the only notable peak of volatility in traditional currencies is the dollar-pound quotation in mid-2016, coinciding with the "Brexit" referendum. Over the whole analyzed perior, the volatility levels of bitcoin was significantly higher than the three exchange rates (in they were plotted at the same scale, the exchange rates volatility would resemble a straight line), but at the same time significantly lower than the other cryptocurrencies. The heterogeneity between different cryptocurrencies motivated the inclusion of the volatility estimation of different cryptocurrencies so that the predictive performance of both traditional econometric models and the machine learning approach can be observed in different cryptocurrencies, generating further evidences of their quality in various contexts.

Since the cryptocurrencies market, notably bitcoin, operates without a supervising organization or entity to ensure it's value or conduct the transactions, as we see in stock exchanges institutions, the economic agents may find difficult to predict this asset, since there is no history or methodology established in the academic or business environment. The construction of a machine, capable to forecast the risk variable, interpreted as volatility, represents an improvement in the studies and business operation using this kind of currency. Furthermore, it may be the first step in the construction of pricing models for the cryptocurrencies and possibles derivatives of it. Therefore, the application of machine learning methods seems very attractive to capture underlying patterns regarding those issues, thus providing a more comprehensive and accurate view of this new agenda.

Analyzing the present techniques and concepts used in the scientific literature for cryptocurrencies estimation, there are few studies that use the Machine Learning approach in forecasting its volatility, even with well known methodologies such as the GARCH model (LI; WANG, 2016). In special, most applications of machine learning methods in finance compare predictive performance based on error metrics like directional accuracy, mean squared error and mean absolute error, instead of using a statistical test as criteria: this procedure is recurrent in papers that apply state-of-art machine learning methods in finance, as seen in Evans, Pappas and Xhafa (2013), Sermpinis et al. (2015) and Shen, Chao and Zhao (2015) . Concerning the superiority of the SVR model over GARCH, many recent works (GAVRISHCHAKA; GANGULI, 2003; GAVRISHCHAKA; BANERJEE, 2006; CHEN; JEONG; HÄRDLE, 2008; PREMANODE; TOUMAZOU, 2013; SANTAMARÍA-BONFIL; FRAUSTO-SOLÍS; VÁZQUEZ-RODARTE, 2015) present favorable evidences towards SVR's superiority, but without enunciating it based on a stronger statistical criteria. Bearing in mind those issues related to financial aspects of cryptocurrencies, this paper combined the machine learning approach with volatility forecasting, splitting the analysis into datasets of high and low frequencies and evaluating the predictive performance of the models using hypothesis tests in order to check in what extent SVR's superiority really holds.

3.3 Method

3.3.1 Volatility estimation

Within the field of financial study, the learning and analysis of financial time series has risen much interest among researchers until today. Technology advance allowed the expansion of financial market and, consequently, of trading operations, which increased the availability of information, quantity of transactions carried out during the day and, mainly, the track of real time assets prices. Regarding financial series analysis, volatility forecasting bears a huge importance, as it has decisive impacts on risk management and derivatives pricing. One of the main stylized facts of this literature states that financial series' conditional variance is typically non-constant. According to Deboeck (1994), Abu-Mostafa and Atiya (1996) and Cao and Tay (2003) these series presents dynamisms and the distribution also shows great variations over time, without exhibiting an apparent and constant pattern in its' disposal. In the financial time series analysis, one can divide the data according to their frequency over time: (1) monthly frequency are usually classified as low frequency. They present a more extensive analysis on macroeconomic variables and analysis of resource allocation and on investment evaluation (EASLEY; PRADO; O'HARA, 2012); (2) data with appearance in minutes or seconds are commonly classified as high frequency. They are strongly influenced by recent events or the availability of

market information, as discussed by Reboredo, Matías and Garcia-Rubio (2012); (3) daily frequency data is the usual periodicity for financial data analysis, such as stock market prediction and volatility estimation, but in recent years, the literature has been moving towards increasingly higher time frequencies or even a volume based paradigm, so that the daily data is used as a low frequency baseline to which the higher frequency data are compared (EASLEY; PRADO; O'HARA, 2012). In this paper, we considered daily data as low frequency and used the hourly periodicity as high frequency.

Due to the importance of fresh news, regarding assets price, Andersen and Bollerslev (1998) explain that financial series present an extremely volatility behavior since they incorporate expectations and reactions of economic agents in the face of events. Currently, market asset volatility forecast and estimation are highly relevant in the composition of derivatives prices, in the portfolio risk analysis and in the investment risk analysis itself. So, the development of methods that help decision making arouses great interest among investors.

Particularly, the high frequency data analysis has caught the attention of many scholars and financial market agents, given the sharp increase in worldwide financial transaction flows, which makes high frequency trading a relevant paradigm for nowadays finance, as discussed in Easley, Prado and O'Hara (2012). With respect to volatility estimation, studies like Li and Wang (2016) indicate that exchange rates and cryptocurrencies' intra-day volatility tend to be very high, motivating its analysis using high frequency data, as seen in Çelik and Ergin (2014) and Baruník and Křehlík (2016).

The standard model used by the academy for volatility estimation is the GARCH model (BOLLERSLEV, 1986), which was derived from the ARCH model (ENGLE, 1982). The GARCH model main advantage is its ability to generalize an ARCH(∞), making it a parsimonious and efficient model to deal with many typical behaviors of financial time series volatility, as highlighted by Marcucci et al. (2005). Furthermore, this model is broadly studied and used by financial analysts, for instance Hansen and Lunde (2005) compared 330 ARCH-type models in terms of their ability to describe the conditional variance using Diebold-Mariano (DIEBOLD; MARIANO, 1995) predictive accuracy test and found no evidence that a GARCH(1,1) was outperformed by more sophisticated models in their exchange rates analysis, but they concluded that GARCH(1,1) was inferior to models that can accommodate a leverage effect in the stock's market analysis.

Recently, other techniques to predict volatility have been discussed, a strong and consistent method used is the Support Vector Regression (SVR), which covers the nonlinearity and dynamic characteristic of the financial series. Gavrishchaka and Ganguli (2003), Gavrishchaka and Banerjee (2006) and Chen, Jeong and Härdle (2008) have already presented empirical results regarding the efficiency and superior predictability of volatility using SVR when compared to the GARCH benchmark and other techniques,

such as neural networks and technical analysis (BARUNÍK; KŘEHLÍK, 2016).

In the last years, there has been a great interest in using traditional methodologies, such as the GARCH model and new ones, like SVR and neural networks, to estimate new assets volatility in the stock or commodities market. As pointed out in Hsu et al. (2016), the machine learning approach has been consistently outperforming traditional econometric models in many research fields inside the finance literature, making such class of methods hugely popular in recent works.

3.3.2 GARCH models

Given P_t the observed price at time $t = 1, \ldots, T$, the GARCH(1,1) (Generalized Auto-regressive Conditional Heteroskedasticity) model can be summarized as follows:

$$r_t = \mu_t + \epsilon_t \tag{3.1}$$

where $r_t = log(\frac{P_t}{P_{t-1}})$ is the log return and ϵ_t is a stochastic term with zero mean. The mean equation of the return in equation **3.1** is defined by an AR(1) model:

$$\mu_t = \gamma_0 + \gamma_1 r_{t-1} \tag{3.2}$$

and the volatility equation is given by:

$$h_t = \alpha_0 + \alpha_1 \epsilon_{t-1}^2 + \beta_1 h_{t-1} \tag{3.3}$$

such that $\mathbb{V}(\epsilon_t) = h_t$.

Since the actual volatility is not directly obtained, an *ex-post* proxy volatility is needed to estimate the volatility through SVR. Following Chen, Härdle and Jeong (2010) and Bezerra and Albuquerque (2017), we defined the proxy volatility as:

$$\tilde{h}_t = (r_t - \bar{r})^2 \tag{3.4}$$

where $\bar{r} = \sum_{t=1}^{T} \frac{r_t}{N}$ is the arithmetic mean of the log returns over the N periods of the in-sample data.

The GARCH (1,1) is one of the main models regarding volatility estimation in finance, given its easy estimation, low number of parameters and ability to capture volatility clusters and conditional variance's non-constant behavior, (HANSEN; LUNDE, 2005). Through a visual analysis of the volatility graphs displayed in figures **2** to **5**, the cryptocurrencies volatilities exhibit a clustering behavior in both low and high frequencies, similarly to the traditional currencies exchange rate pattern, in a much higher magnitude

nevertheless, such that we can conclude that the volatility clustering stylized fact seems to hold for the dataset we analyzed, which in turn justifies the use of GARCH models to estimate their volatility.

Even when ϵ_t is assumed to be normally distributed, GARCH presents a fat-tailed behavior in comparison to the Gaussian distribution, even though still not quite incorporating the financial data's stylized facts (CONT, 2001). Thus, it is common to assume that ϵ_t follows non-Gaussian distributions, in order to fit better to the financial data. In this paper, we estimated the GARCH (1,1) assuming three conditional distributions for ϵ_t: Normal, Student's t and Skewed Student's t. Despite the fact that literature proposed a wide variety of distributions to be assumed in the GARCH model, authors like Sun and Zhou (2014) argue that the Student's t is enough to give a good fit to the financial data's heavy tail behavior, while innovations concerning GARCH conditional distributions does not seem robust.

As discussed by studies like Awartani and Corradi (2005), the asymmetric volatility (also known as the "leverage effect") is a well know stylized fact in financial markets: typically, a negative shock in t tend to make a higher impact on the volatility at $t+1$ than positive shocks. Thus, even though the literature presented many positive evidences towards GARCH (1,1) model over a great number of conditional heteroskedasticity models (HANSEN; LUNDE, 2005), studies like Awartani and Corradi (2005),Wang (2009) and Laurent, Rombouts and Violante (2012) present evidences that GARCH extensions that incorporate the effects of asymmetry in financial series' volatility, such as EGARCH and GJR-GARCH, yielded smaller out-of-sample prediction errors in comparison to the standard GARCH. Thus, we incorporated EGARCH (1,1) and GJR-GARCH (1,1) as benchmarks as well.

The volatility equation for EGARCH (1,1) (NELSON, 1991) is given by:

$$\log(h_t) = \alpha_0 + g(z_{t-1}) + \beta_1 \log(h_{t-1}) \tag{3.5}$$

where $\epsilon_t = h_t z_t$ and $g(z_t) = \alpha_1 z_t + \gamma[|z_t| - \mathbb{E}(|z_t|)]$ and $z_t \sim N(0,1)$. And the volatility equation of GJR-GARCH (1,1) (GLOSTEN; JAGANNATHAN; RUNKLE, 1993) is given by:

$$\begin{aligned} h_t &= \alpha_0 + (\alpha_1 + \gamma_1 I_{t-1})\epsilon_{t-1}^2 + \beta_1 h_{t-1} \\ I_t &= \begin{cases} 0, & \epsilon_t \geq 0, \\ 1, & \epsilon_t < 0. \end{cases} \end{aligned} \tag{3.6}$$

3.3.3 Support Vector Regression

Despite its popularity, GARCH (1,1) still considers linear functional forms in its estimation, which motivates the introduction of nonlinear structural forms. That is one of the main contributions of machine learning and kernel methods, as many studies showed (LI; SUOHAI, 2013; SHEN; CHAO; ZHAO, 2015) that the introduction of nonlinear interactions can significantly boost the explanatory power and the forecasting ability of many models applied to financial contexts, including volatility estimation (CHEN; HÄRDLE; JEONG, 2010; PREMANODE; TOUMAZOU, 2013; SANTAMARÍA-BONFIL; FRAUSTO-SOLÍS; VÁZQUEZ-RODARTE, 2015). Regarding high frequency forecasting, that issue is also noted (SANTOS; COSTA; COELHO, 2007).

Therefore, we used Support Vector Regression to estimate of the GARCH's mean and volatility equations – described in equations **3.2** and **3.3**. Instead of using a standard linear regression, we introduced nonlinearities, in order to provide a better fit to the data. Concerning the high volatility of bitcoin data – specially in high frequency transactions - this approach seems particularly attractive.

The Support Vector Regression (SVR) (VAPNIK, 1995; DRUCKER et al., 1997) is a regression model that aims to find a decision function which is the best approximation of a set of observations, bearing in mind the middle ground between a good power of generalization and an overall stable behavior, in order to make good out-of-sample inferences. Associated with these two desirable features, there are two corresponding problems in regression models, constituting the so called "bias-variance dilemma". To perform the regularization of the decision function, two parameters are added: a band of tolerance δ[2], to avoid over-fitting; and a penalty C to the objective function, for points that lie outside this confidence interval for an amount $\boldsymbol{\xi} > 0$.

Therefore, the predicted values $f(\mathbf{x_i})$, such that $|y_i - f(\mathbf{x_i})| \leq \delta$ and $f(.)$ is the SVR decision function, are considered to be statistically equal to y. The SVR defined from the addition of these two parameters is known as ε-SVR. The loss function implied in the construction of the ε-SVR is the ε-insensitive loss function (VAPNIK, 1995), $L_\varepsilon[y_i, f(\mathbf{x_i})]$, given by:

$$L_\varepsilon[y_i, f(\mathbf{x_i})] = \begin{cases} |y_i - f(\mathbf{x_i})| - \delta, & |y_i - f(\mathbf{x_i})| > \delta, \\ 0, & |y_i - f(\mathbf{x_i})| \leq \delta. \end{cases} \tag{3.7}$$

It is worth noting that the ε-insensitive loss function is not the only possible way to define penalties for SVR; extensions that include different penalty structures includes the ν-SVR (CHANG; LIN, 2002). The ε-SVR formulation was chosen for this paper because

[2] ε is the usual symbol used in SVR models for the confidence band. In this paper, we changed ε to δ to avoid ambiguities with the GARCH model error term ϵ

it is the most commonly used form in the finance forecasting literature, and requires lesser computational time to perform the optimization.

In order to introduce nonlinear interactions in the regression estimation, a mapping φ is applied, such that the objective function to be optimized for ε-SVR is formulated as follows:

$$\begin{aligned} \text{Minimize :} \quad & \frac{1}{2}\mathbf{w}^\top\mathbf{w} + C\boldsymbol{\xi}^\top\mathbf{1} + C{\boldsymbol{\xi}^*}^\top\mathbf{1} \\ \text{Subject to :} \quad & \Phi\mathbf{w} + w_0 - \mathbf{y} \leq \delta\mathbf{1} + \boldsymbol{\xi} \\ & \mathbf{y} - \Phi\mathbf{w} - w_0 \leq \delta\mathbf{1} + \boldsymbol{\xi}^* \\ \text{with :} \quad & \boldsymbol{\xi}, \boldsymbol{\xi}^* \geq 0 \end{aligned} \tag{3.8}$$

where Φ is a $T \times q$ matrix created by the Feature Space, i.e., the original explanatory variables $\mathbf{X}_{(T\times p)}$ is mapped through the $\varphi(\mathbf{x})$ function, $\mathbf{w}$ is a vector of parameters to be estimated, C and δ are hyper-parameters and $\boldsymbol{\xi}, \boldsymbol{\xi}^*$ are slack variables in the Quadratic Programming Problem.

In other words, $\mathbf{w}_{(q\times 1)}$ is the vector of the angular coefficients of the decision hyperplane in $\mathbb{R}^q$; $w_0 \in \mathbb{R}$ is the linear coefficient (intercept) of the decision hyperplane in $\mathbb{R}^q$; $\Phi_{(T\times q)}$ is the augmented matrix of observations, after the original data being transformed by φ; $\mathbf{y}_{(T\times 1)}$ is the vector that provides the dependent variable values of the observed points; $C \in \mathbb{R}$ is the cost of error; $\delta > 0$ is the tolerance band that defines the confidence interval for which there is no penalty; ${\boldsymbol{\xi}^*}_{(T\times 1)}$ is the vector concerning points above the tolerance band; and $\boldsymbol{\xi}_{(T\times 1)}$ is the vector concerning points below the tolerance band.

After some algebraic manipulations, it can be shown that the decision function of ε-SVR can be written as:

$$f(\mathbf{x_i}) = \mathbf{w}^\top\varphi(\mathbf{x}) - w_0 = \sum_{t=1}^{T} \kappa(\mathbf{x_i}, \mathbf{x_j})(\lambda_j^* - \lambda_j) - w_0 \tag{3.9}$$

where $\kappa(\mathbf{x_i}, \mathbf{x_j}) = \varphi(\mathbf{x_i}) \cdot \varphi(\mathbf{x_j}) \in \mathbb{R}, i, j = 1, 2, 3 \ldots, T$ is the kernel function. Since φ transforms the original data to a higher dimension, which can even be infinite, the use of the kernel function prevents the need to explicitly compute the functional form of $\varphi(\mathbf{x})$; instead, κ computes the inner product of φ, a term that appears in SVR's dual formulation (DRUCKER et al., 1997). This is known as the *kernel trick*. In this paper, we used the Gaussian Kernel as κ, whose expression is given by:

$$\kappa(\mathbf{x_i}, \mathbf{x_j}) = exp\left(-\frac{||\mathbf{x_i} - \mathbf{x_j}||^2}{2\sigma^2}\right), \sigma > 0 \tag{3.10}$$

The Gaussian Kernel is the most widely used Kernel in machine learning literature. It enjoys huge popularity in various knowledge fields since this function is able to induce an infinite dimensional feature space while depending on only one scattering parameter σ.

3.3.4 SVR-GARCH

The SVR-GARCH is the joining result of the GARCH model structure and the nonlinearities introduced by the kernel function via SVR. Santamaría-Bonfil, Frausto-Solís and Vázquez-Rodarte (2015) presented empirical evidences that SVR-GARCH managed to outperform standard GARCH's predictions, showing better ability to approximate the nonlinear behavior of financial data and stylized facts, such as heavy tails and volatility clusters. The specification of SVR-GARCH (1,1) is the same of the conventional GARCH (1,1), but the mean and volatility equations were estimated via SVR, such that:

$$r_t = f_m(r_{t-1}) + \epsilon_t \tag{3.11}$$

$$h_t = f_v(h_{t-1}, \epsilon_{t-1}^2) \tag{3.12}$$

where $f_m(.)$ is the SVR decision function for the mean equation **3.1**, and $f_v(.)$ is the SVR decision function for the volatility equation **3.3**.

Depending of which parameters δ, C and σ are set to the SVR formulation, a different decision function is obtained. In order to decide the "better" combination of parameters, we did a grid search for those three parameters for both mean and volatility equations, and evaluated the Root Mean Square Error (RMSE) of each decision function. The model is first estimated with a subset of the data (known as training dataset) and then the estimated model is used to forecast both mean and volatility in another subset (validation dataset). The combination that minimizes the RMSE for the validation dataset was chosen as the best one. This combination of parameters was applied to yield the forecasts of the out-of-sample data, known as the test set. The search intervals for each parameter of the SVR-GARCH are displayed in Table **1**.

Parameter	Search interval
δ	$[0.05, 0.1, \ldots, 0.95, 1]$
C	$[0.5, 1, \ldots, 4.5, 5]$
σ	$[0.05, 0.1, \ldots, 1.95, 2]$

Table 1 – Search intervals used for the parameters' training

3.4 Empirical analysis

For the empirical test, we used data between January 4th 2016 and July 31st 2017 of three cryptocurrencies: bitcoin, ethereum and dash market price (in US dollars); and three traditional currencies: euro, british pound and japanese yen (in US dollars). The data was collected from Altcoin Charts (http://alt19.com) and Forex Historical Data (http://fxhistoricaldata.com). Since FOREX data are not available for weekends, we collected only the weekdays in all variables to assure that the models were fitted in the same days. We used the daily basis for low frequency analysis (411 observations) and the hour periodicity for high frequency estimation (9742 observations).

Both databases were partitioned into three mutually exclusive datasets: training set, validation set and test set. The purpose of this segmentation is to allow the machine learning algorithm to test its predictive performance on data that were not priorly used, in order to better evaluate the real explanatory power of the found decision function when dealing with new data. The training and validation sets constitute the in-sample data for GARCH models and the test set constitute the out-of-sample data in which the predictions were made. The horizon of the forecasts was one step ahead (one day for low frequency and one hour for high frequency data).

For low frequency data, we allocated 10 months for training (January 2016 to October 2016 - 216 days), 4 months for validation (November 2016 to February 2017 - 84 days) and 5 months for test (March 2017 to July 2017 - 119 days).

In order to verify the robustness of SVR-GARCH model's predictive performance over time, we split the high frequency dataset with a moving window over the whole period, defining smaller time periods with 9 months each. In each period, we allocated the first 5 months for training, the next 2 months for validation and the last 2 months for test. For the following time intervals, the time periods were shifted two months forward until the end of the dataset extension, totaling six time periods of hourly data. Inside the subsets, we did not apply rolling windows for the estimations.

Therefore, the time periods for high frequency data were defined as follows:

- Period 1: January 4th 2016 to September 30th 2016.
- Period 2: March 1st 2016 to November 30th 2016.
- Period 3: May 2nd 2016 to January 31st 2017.
- Period 4: July 1st 2016 to March 31st 2017.
- Period 5: September 1st 2016 to May 31st 2017.
- Period 6: November 1st 2016 to July 31st 2017.

Firstly, the optimization of the SVR algorithm was applied to each training set, by performing a grid search for each one of the associated parameters for both mean and volatility equations in low and high frequencies. The search ranges for the parameters δ, C and σ are listed in Table **1**.

Based on each combination of parameters applied to the training set, the accuracy of each optimal obtained decision function was checked for the validation set using the error metric RMSE, defined as:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{T} (\tilde{h}_t - \widehat{h}_t)^2}{T}} \tag{3.13}$$

where $\tilde{h}_t$ is the proxy volatility as defined in **3.4** and $\tilde{h}_t$ is the predicted volatility.

Each decision function obtained in the training set was fed with data from the validation set to compute the prediction of the dependent variable for these data. This forecast was then confronted with the actual values observed in the validation set, then the RMSE between predicted and observed values was calculated. Repeating the process for every parameter combination, the optimal combination is the one that minimizes the RMSE associated with its prediction. For the GARCH models, the training and validation sets were jointly used as in-sample data to estimate the respective coefficients.

Subsequently, the optimal parameters were applied to fit the model for the test set, and then compared with the results generated by GARCH models, obtaining the one-step ahead volatility estimation for the time periods of each test set, totaling seven sets of forecasts (one for daily data and six for hourly data). For this step, we considered the error metrics RMSE, defined in equation **3.13**; and MAE (mean absolute error), whose expression is given by:

$$MAE = \frac{\sum_{t=1}^{T} |\tilde{h}_t - \widehat{h}_t|}{T} \tag{3.14}$$

Finally, in order to check the statistical significance of SVR-GARCH's superiority over GARCH models, we applied Diebold and Mariano (1995) predictive accuracy test for the nine GARCH models in both low and high frequencies, using SVR-GARCH as benchmark. Additionally, we applied the Model Confidence Test (HANSEN; LUNDE; NASON, 2011) for each set of forecasts to further investigate whether the machine learning based approach yielded significantly better results. The description of both tests are displayed in appendices **A** and **B**.

Training set: January 4th 2016 to October 31st 2016
Validation set: November 1st 2016 to February 28th 2017
Test set: March 1st 2017 to July 31st 2017

Model	Bitcoin		Ethereum		Dash	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
Normal GARCH (1,1)	0.11262520	0.11025030	0.26447120	0.24894240	0.19964301	0.16836890
Student's *t* GARCH (1,1)	0.15802830	0.15560480	0.33843840	0.33294630	0.20517540	0.18140470
Skewed Student's *t* GARCH (1,1)	0.15603930	0.15356490	0.60801570	0.59901870	0.20540960	0.18172430
Normal EGARCH (1,1)	0.14851058	0.06021389	0.24741551	0.12332978	0.18845499	0.10440545
Student's *t* EGARCH (1,1)	0.15294444	0.13761023	0.29789732	0.27103321	0.19264046	0.11003185
Skewed Student's *t* EGARCH (1,1)	0.15437027	0.13844548	0.86978330	0.83615954	0.19297362	0.11041533
Normal GJR-GARCH (1,1)	0.15878742	0.06073869	0.24829644	0.18500229	0.18631002	0.10449338
Student's *t* GJR-GARCH (1,1)	0.15060326	0.13426739	0.29438251	0.26876020	0.19277590	0.11011040
Skewed Student's *t* GJR-GARCH (1,1)	0.14354459	0.17748075	0.61748915	0.58003697	0.19363469	0.11046644
SVR-GARCH (1,1)	0.03133926	0.01315455	0.20422370	0.08627904	0.15282070	0.04538759
Model	Euro		British Pound		Japanese Yen	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
Normal GARCH (1,1)	0.01163661	0.01162828	0.00936268	0.00936006	0.01028482	0.01028177
Student's *t* GARCH (1,1)	0.02416139	0.02374070	0.01008274	0.01007267	0.01103492	0.01102153
Skewed Student's *t* GARCH (1,1)	0.02411368	0.02369469	0.01009985	0.01008939	0.01098031	0.01096716
Normal EGARCH (1,1)	0.00557848	0.00473396	0.00688289	0.00539079	0.00724395	0.00592201
Student's *t* EGARCH (1,1)	0.00556659	0.00467683	0.00691821	0.00542716	0.00745208	0.00608476
Skewed Student's *t* EGARCH (1,1)	0.00585727	0.00498910	0.00691028	0.00541879	0.00766734	0.00631347
Normal GJR-GARCH (1,1)	0.00580383	0.00499980	0.00680241	0.00537477	0.00716407	0.00590922
Student's *t* GJR-GARCH (1,1)	0.00580788	0.00500314	0.00674019	0.00531741	0.00715457	0.00589986
Skewed Student's *t* GJR-GARCH (1,1)	0.00580250	0.00498450	0.00675847	0.00533939	0.00715230	0.00589715
SVR-GARCH (1,1)	0.00030316	0.00011757	0.00023233	0.00008382	0.00022602	0.00014921

The Hansen, Lunde and Nason (2011)'s superior set models are highlighted in gray

Table 2 – Forecasting performance for low frequency test set data

Training set: January 4th 2016 to May 31st 2016
Validation set: June 1st 2016 to July 31st 2016
Test set: August 1st 2016 to September 30th 2016

Model	Bitcoin		Ethereum		Dash	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
Normal GARCH (1,1)	0.00759729	0.00569709	0.01080670	0.00972066	0.01514854	0.01404126
Student's *t* GARCH (1,1)	0.00817163	0.00600172	0.01155509	0.00999820	0.01446214	0.01363025
Skewed Student's *t* GARCH (1,1)	0.00797165	0.00580175	0.01156629	0.01000044	0.01442460	0.01359423
Normal EGARCH (1,1)	0.00660961	0.00546262	0.01066027	0.00972147	0.01490163	0.01395768
Student's *t* EGARCH (1,1)	0.00560289	0.00486634	0.01124901	0.00997496	0.01427962	0.01358828
Skewed Student's *t* EGARCH (1,1)	0.00563103	0.00487856	0.01125544	0.00998051	0.01428608	0.01358456
Normal GJR-GARCH (1,1)	0.00774527	0.00570928	0.01082471	0.00968486	0.01529941	0.01407232
Student's *t* GJR-GARCH (1,1)	0.00564167	0.00491907	0.01158647	0.00998025	0.01457111	0.01367225
Skewed Student's *t* GJR-GARCH (1,1)	0.00555230	0.00483404	0.01159366	0.00998519	0.01456874	0.01366403
SVR-GARCH (1,1)	0.00059805	0.00026484	0.00068952	0.00021251	0.00092300	0.00020908
Model	Euro		British Pound		Japanese Yen	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
Normal GARCH (1,1)	0.00082456	0.00082383	0.00118071	0.00118059	0.00132147	0.00131698
Student's *t* GARCH (1,1)	0.00080980	0.00080745	0.00130299	0.00129675	0.00138082	0.00136957
Skewed Student's *t* GARCH (1,1)	0.00089016	0.00088014	0.00126084	0.00125772	0.00129791	0.00128906
Normal EGARCH (1,1)	0.00081491	0.00081073	0.00119775	0.00119498	0.00132418	0.00130711
Student's *t* EGARCH (1,1)	0.00089333	0.00081018	0.00130253	0.00123891	0.00134704	0.00133007
Skewed Student's *t* EGARCH (1,1)	0.00089234	0.00081048	0.00130001	0.00123844	0.00134781	0.00133081
Normal GJR-GARCH (1,1)	0.00084235	0.00083376	0.00124946	0.00123813	0.00137471	0.00135760
Student's *t* GJR-GARCH (1,1)	0.00077467	0.00076647	0.00119304	0.00118418	0.00125635	0.00124371
Skewed Student's *t* GJR-GARCH (1,1)	0.00077294	0.00076013	0.00128225	0.00127839	0.00131223	0.00130378
SVR-GARCH (1,1)	0.00024853	0.00009478	0.00072465	0.00050004	0.00005872	0.00002306

The Hansen, Lunde and Nason (2011)'s superior set models are highlighted in gray

Table 3 – Forecasting performance for high frequency test set data: Period 1

Training set: March 1st 2016 to July 31st 2016
Validation set: August 1st 2016 to September 30th 2016
Test set: October 1st 2016 to November 30th 2016

Model	Bitcoin		Ethereum		Dash	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
Normal GARCH (1,1)	0.00460814	0.00440171	0.00775727	0.00770857	0.01163329	0.01151671
Student's t GARCH (1,1)	0.00479175	0.00467229	0.00782142	0.00765811	0.01191285	0.01171240
Skewed Student's t GARCH (1,1)	0.00414653	0.00392619	0.00780785	0.00764424	0.01191912	0.01171801
Normal EGARCH (1,1)	0.00449658	0.00436728	0.00782352	0.00775327	0.01162451	0.01149187
Student's t EGARCH (1,1)	0.00430255	0.00418052	0.00770778	0.00758103	0.01190195	0.01170119
Skewed Student's t EGARCH (1,1)	0.00431631	0.00419205	0.00770736	0.00758063	0.01191014	0.01170721
Normal GJR-GARCH (1,1)	0.00465698	0.00438188	0.00483759	0.00371794	0.01165126	0.01152582
Student's t GJR-GARCH (1,1)	0.00441468	0.00422006	0.00762143	0.00748573	0.01193339	0.01171407
Skewed Student's t GJR-GARCH (1,1)	0.00441076	0.00421270	0.00779723	0.00763222	0.01194239	0.01172274
SVR-GARCH (1,1)	0.00077868	0.00020442	0.00177067	0.00056101	0.00307538	0.00127483
Model	Euro		British Pound		Japanese Yen	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
Normal GARCH (1,1)	0.00111576	0.00101530	0.00161297	0.00155571	0.00174503	0.00147923
Student's t GARCH (1,1)	0.00120193	0.00106713	0.00167885	0.00162099	0.00153819	0.00141659
Skewed Student's t GARCH (1,1)	0.00115309	0.00103867	0.00169058	0.00162774	0.00154086	0.00142141
Normal EGARCH (1,1)	0.00109249	0.00102655	0.00157519	0.00153497	0.00149050	0.00139378
Student's t EGARCH (1,1)	0.00115206	0.00107847	0.00199462	0.00171057	0.00152148	0.00141757
Skewed Student's t EGARCH (1,1)	0.00115383	0.00108003	0.00199594	0.00171036	0.00152152	0.00141599
Normal GJR-GARCH (1,1)	0.00114574	0.00104284	0.00013351	0.00000491	0.00154525	0.00139834
Student's t GJR-GARCH (1,1)	0.00117087	0.00108371	0.00165860	0.00159177	0.00159770	0.00145711
Skewed Student's t GJR-GARCH (1,1)	0.00114196	0.00105452	0.00164845	0.00159493	0.00007137	0.00008873
SVR-GARCH (1,1)	0.00050881	0.00015161	0.00011788	0.00002368	0.00010146	0.00005867

The Hansen, Lunde and Nason (2011)'s superior set models are highlighted in gray

Table 4 – Forecasting performance for high frequency test set data: Period 2

Training set: May 1st 2016 to September 30th 2016
Validation set: October 1st 2016 to November 30th 2016
Test set: December 1st 2016 to January 31st 2017

Model	Bitcoin		Ethereum		Dash	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
Normal GARCH (1,1)	0.00913742	0.00749972	0.01270315	0.01127455	0.01415568	0.01325402
Student's t GARCH (1,1)	0.00925140	0.00722405	0.01294731	0.01139057	0.01564308	0.01437991
Skewed Student's t GARCH (1,1)	0.00925199	0.00722519	0.01294729	0.01138956	0.01564466	0.01437814
Normal EGARCH (1,1)	0.00955980	0.00785587	0.01294379	0.01147051	0.01340430	0.01302949
Student's t EGARCH (1,1)	0.01113715	0.00888592	0.01284171	0.01138972	0.01519257	0.01426543
Skewed Student's t EGARCH (1,1)	0.01119126	0.00892786	0.01284247	0.01139012	0.01538706	0.01444303
Normal GJR-GARCH (1,1)	0.00869412	0.00727078	0.01277006	0.01130760	0.01392639	0.01348153
Student's t GJR-GARCH (1,1)	0.00931768	0.00722120	0.01299166	0.01139946	0.01539582	0.01413736
Skewed Student's t GJR-GARCH (1,1)	0.00932389	0.00722400	0.01299131	0.01139830	0.01544307	0.01416606
SVR-GARCH (1,1)	0.00042599	0.00008641	0.00050119	0.00030067	0.00076065	0.00020646
Model	Euro		British Pound		Japanese Yen	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
Normal GARCH (1,1)	0.00133953	0.00131835	0.00151323	0.00148351	0.00161222	0.00158059
Student's t GARCH (1,1)	0.00136766	0.00129345	0.00157570	0.00156285	0.00159663	0.00156250
Skewed Student's t GARCH (1,1)	0.00154868	0.00146029	0.00162702	0.00160678	0.00157617	0.00154267
Normal EGARCH (1,1)	0.00134860	0.00132980	0.00158497	0.00144133	0.00159121	0.00155684
Student's t EGARCH (1,1)	0.00153668	0.00144513	0.00157096	0.00150291	0.00162717	0.00159317
Skewed Student's t EGARCH (1,1)	0.00156194	0.00146674	0.00156715	0.00149893	0.00163267	0.00159789
Normal GJR-GARCH (1,1)	0.00134752	0.00133283	0.00154829	0.00149573	0.00163351	0.00159968
Student's t GJR-GARCH (1,1)	0.00135133	0.00131171	0.00155834	0.00154596	0.00160787	0.00156749
Skewed Student's t GJR-GARCH (1,1)	0.00135251	0.00130687	0.00155696	0.00153681	0.00157856	0.00153916
SVR-GARCH (1,1)	0.00070290	0.00020679	0.00080172	0.00022069	0.00062456	0.00022606

The Hansen, Lunde and Nason (2011)'s superior set models are highlighted in gray

Table 5 – Forecasting performance for high frequency test set data: Period 3

Training set: July 1st 2016 to November 30th 2016
Validation set: December 1st 2016 to January 31st 2017
Test set: February 1st 2017 to March 31st 2017

Model	Bitcoin		Ethereum		Dash	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
Normal GARCH (1,1)	0.01028286	0.00902283	0.01793080	0.01505344	0.02825603	0.02666399
Student's t GARCH (1,1)	0.01032863	0.00883706	0.01710276	0.01338831	0.02649039	0.02170007
Skewed Student's t GARCH (1,1)	0.01033805	0.00884552	0.01710271	0.01338835	0.02649035	0.02170456
Normal EGARCH (1,1)	0.01016503	0.00900445	0.01862030	0.01599543	0.02905840	0.02581053
Student's t EGARCH (1,1)	0.01036353	0.00897735	0.01890077	0.01521900	0.03421714	0.02878764
Skewed Student's t EGARCH (1,1)	0.01045986	0.00898360	0.01890052	0.01521885	0.03429000	0.02884370
Normal GJR-GARCH (1,1)	0.01034703	0.00900930	0.01960318	0.01560942	0.02827459	0.02584707
Student's t GJR-GARCH (1,1)	0.01084610	0.00903140	0.01749649	0.01346106	0.02661699	0.02150698
Skewed Student's t GJR-GARCH (1,1)	0.01106700	0.00915978	0.01749658	0.01346125	0.02661630	0.02150540
SVR-GARCH (1,1)	0.00063104	0.00017862	0.00252631	0.00099954	0.00597564	0.00120334
Model	Euro		British Pound		Japanese Yen	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
Normal GARCH (1,1)	0.00092297	0.00091336	0.00114754	0.00114037	0.00116164	0.00115678
Student's t GARCH (1,1)	0.00096412	0.00095584	0.00133873	0.00133610	0.00116044	0.00114760
Skewed Student's t GARCH (1,1)	0.00090031	0.00089295	0.00117397	0.00116803	0.00116685	0.00115889
Normal EGARCH (1,1)	0.00089818	0.00088116	0.00116322	0.00111639	0.00114097	0.00113461
Student's t EGARCH (1,1)	0.00094176	0.00092113	0.00131765	0.00127703	0.00114329	0.00113341
Skewed Student's t EGARCH (1,1)	0.00094128	0.00092054	0.00131757	0.00127717	0.00114347	0.00113358
Normal GJR-GARCH (1,1)	0.00096640	0.00095289	0.00117486	0.00115431	0.00117479	0.00116757
Student's t GJR-GARCH (1,1)	0.00088538	0.00087678	0.00125185	0.00124700	0.00113320	0.00112218
Skewed Student's t GJR-GARCH (1,1)	0.00093148	0.00092358	0.00120794	0.00120456	0.00110560	0.00109304
SVR-GARCH (1,1)	0.00002209	0.00001039	0.00043386	0.00014066	0.00041692	0.00016176

The Hansen, Lunde and Nason (2011)'s superior set models are highlighted in gray

Table 6 – Forecasting performance for high frequency test set data: Period 4

Training set: September 1st 2016 to January 31st 2017
Validation set: February 1st 2017 to March 31st 2017
Test set: April 1st 2017 to May 31st 2017

Model	Bitcoin		Ethereum		Dash	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
Normal GARCH (1,1)	0.01140054	0.00964473	0.01968106	0.01760813	0.02057793	0.01910974
Student's t GARCH (1,1)	0.01148559	0.00971204	0.01873198	0.01621085	0.02052849	0.01852332
Skewed Student's t GARCH (1,1)	0.01148641	0.00971607	0.01872158	0.01622386	0.02049204	0.01850511
Normal EGARCH (1,1)	0.01166953	0.00979897	0.01852513	0.01699511	0.02003365	0.01874298
Student's t EGARCH (1,1)	0.01137208	0.00976844	0.01990472	0.01741331	0.02128344	0.01936220
Skewed Student's t EGARCH (1,1)	0.01139669	0.00978829	0.01970050	0.01726016	0.02127683	0.01935668
Normal GJR-GARCH (1,1)	0.01176680	0.00973092	0.01984240	0.01763048	0.02076537	0.01916559
Student's t GJR-GARCH (1,1)	0.01174897	0.00977212	0.01873072	0.01608512	0.02041031	0.01828486
Skewed Student's t GJR-GARCH (1,1)	0.01177109	0.00978936	0.01871247	0.01609221	0.02039492	0.01827536
SVR-GARCH (1,1)	0.00050642	0.00021148	0.00173549	0.00124855	0.00157584	0.00118765
Model	Euro		British Pound		Japanese Yen	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
Normal GARCH (1,1)	0.00103368	0.00089070	0.00092294	0.00090855	0.00113294	0.00112792
Student's t GARCH (1,1)	0.00092145	0.00087629	0.00090351	0.00089595	0.00110693	0.00108632
Skewed Student's t GARCH (1,1)	0.00099263	0.00097098	0.00093310	0.00092156	0.00114681	0.00112563
Normal EGARCH (1,1)	0.00145672	0.00090461	0.00091872	0.00089352	0.00111866	0.00110740
Student's t EGARCH (1,1)	0.00107954	0.00087316	0.00098508	0.00094400	0.00111971	0.00111020
Skewed Student's t EGARCH (1,1)	0.00108103	0.00872119	0.00098502	0.00094325	0.00111423	0.00110502
Normal GJR-GARCH (1,1)	0.00093251	0.00089027	0.00091505	0.00090916	0.00115428	0.00113325
Student's t GJR-GARCH (1,1)	0.00091850	0.00089133	0.00091570	0.00089692	0.00072788	0.00040702
Skewed Student's t GJR-GARCH (1,1)	0.00091459	0.00089133	0.00091284	0.00089180	0.00110604	0.00109003
SVR-GARCH (1,1)	0.00008139	0.00000166	0.00018515	0.00008813	0.00050841	0.00013768

The Hansen, Lunde and Nason (2011)'s superior set models are highlighted in gray

Table 7 – Forecasting performance for high frequency test set data: Period 5

Training set: November 1st 2016 to March 31st 2017
Validation set: April 1st 2017 to May 31st 2017
Test set: June 1st 2017 to July 31st 2017

Model	Bitcoin		Ethereum		Dash	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
Normal GARCH (1,1)	0.01341529	0.01289053	0.02267557	0.02171070	0.01540930	0.01384991
Student's *t* GARCH (1,1)	0.01418239	0.01304112	0.02367832	0.02093723	0.02074012	0.01945195
Skewed Student's *t* GARCH (1,1)	0.01416461	0.01302035	0.02366414	0.02091874	0.02088385	0.01954755
Normal EGARCH (1,1)	0.01323332	0.01278953	0.02251908	0.02172502	0.02110245	0.02097898
Student's *t* EGARCH (1,1)	0.01423269	0.01304949	0.02552701	0.02223959	0.01993591	0.01909908
Skewed Student's *t* EGARCH (1,1)	0.01449978	0.01328147	0.02565733	0.02230468	0.02025959	0.01937293
Normal GJR-GARCH (1,1)	0.01359696	0.01288202	0.02281586	0.02170132	0.02119260	0.02107908
Student's *t* GJR-GARCH (1,1)	0.01443705	0.01306728	0.02374665	0.02080738	0.02046007	0.01938376
Skewed Student's *t* GJR-GARCH (1,1)	0.01463602	0.01322887	0.02386151	0.02088308	0.02082588	0.01968665
SVR-GARCH (1,1)	0.00058801	0.00020410	0.00193367	0.00066853	0.00259729	0.00053860
Model	Euro		British Pound		Japanese Yen	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
Normal GARCH (1,1)	0.00090206	0.00089563	0.00113611	0.00111291	0.00101607	0.00101388
Student's *t* GARCH (1,1)	0.00085591	0.00083969	0.00114517	0.00111681	0.00106887	0.00105437
Skewed Student's *t* GARCH (1,1)	0.00087847	0.00086434	0.00109577	0.00106716	0.00108514	0.00107198
Normal EGARCH (1,1)	0.00102635	0.00089134	0.00121050	0.00110969	0.00103235	0.00102242
Student's *t* EGARCH (1,1)	0.00099420	0.00093139	0.00087107	0.00117713	0.00104719	0.00102476
Skewed Student's *t* EGARCH (1,1)	0.00099030	0.00092624	0.00088244	0.00118472	0.00104756	0.00102427
Normal GJR-GARCH (1,1)	0.00093917	0.00091280	0.00123846	0.00000336	0.00102764	0.00102210
Student's *t* GJR-GARCH (1,1)	0.00087293	0.00085482	0.00014394	0.00108982	0.00101661	0.00100684
Skewed Student's *t* GJR-GARCH (1,1)	0.00084019	0.00082184	0.00115170	0.00111520	0.00100928	0.00099567
SVR-GARCH (1,1)	0.00003128	0.00001030	0.00060816	0.00003409	0.00036032	0.00001103

The Hansen, Lunde and Nason (2011)'s superior set models are highlighted in gray

Table 8 – Forecasting performance for high frequency test set data: Period 6

Model	Bitcoin		Ethereum		Dash	
	Test statistic	P-value	Test statistic	P-value	Test statistic	P-value
Normal GARCH (1,1)	3.1076	0.001006*	2.4027	0.008867*	1.3705	0.086480
Student's *t* GARCH (1,1)	3.5682	0.000194*	5.7427	1.56E-07*	1.4245	0.078380
Skewed Student's *t* GARCH (1,1)	3.5106	0.000230*	8.8443	2.21E-16*	1.4316	0.077370
Normal EGARCH (1,1)	2.4695	0.007483*	2.1036	0.018791	1.7968	0.037465
Student's *t* EGARCH (1,1)	3.1232	0.001136*	2.8926	0.002275*	2.2311	0.013787
Skewed Student's *t* EGARCH (1,1)	3.4270	0.000430*	19.6920	2.21E-16*	2.2586	0.012871
Normal GJR-GARCH (1,1)	2.5124	0.006672*	1.5564	0.061071	1.6434	0.051486
Student's *t* GJR-GARCH (1,1)	2.8048	0.002945*	2.7916	0.003064*	2.2425	0.013481
Skewed Student's *t* GJR-GARCH (1,1)	4.5249	0.000007*	14.2278	2.21E-16*	2.2909	0.011872
Model	Period 4		Period 5		Period 6	
	Test statistic	P-value	Test statistic	P-value	Test statistic	P-value
Normal GARCH (1,1)	3.2808	0.000550*	3.2708	0.000692*	3.2790	0.000674*
Student's *t* GARCH (1,1)	2.7967	0.003069*	2.2321	0.013690	5.7872	2.67E-08*
Skewed Student's *t* GARCH (1,1)	2.7077	0.003667*	3.6659	0.000181*	1.7026	0.003977*
Normal EGARCH (1,1)	2.4996	0.006918*	3.2151	0.000849*	3.3026	0.000634*
Student's *t* EGARCH (1,1)	2.4279	0.008354*	3.2834	0.000682*	3.5453	0.000289*
Skewed Student's *t* EGARCH (1,1)	3.2622	0.000720*	3.2699	0.000715*	3.8388	0.000126*
Normal GJR-GARCH (1,1)	3.2095	0.000863*	3.1958	0.000898*	3.2832	0.000685*
Student's *t* GJR-GARCH (1,1)	3.2195	0.000834*	3.1372	0.001087*	3.2660	0.000713*
Skewed Student's *t* GJR-GARCH (1,1)	3.1933	0.000992*	3.1671	0.000980*	3.2615	0.000725*

* denote statistical significance at 1% level

Table 9 – Diebold-Mariano test statistic and p-value for low frequency data

Model	Period 1		Period 2		Period 3	
	Test statistic	P-value	Test statistic	P-value	Test statistic	P-value
Normal GARCH (1,1)	2.0473	0.020400	2.5388	0.005635*	2.8350	0.002338*
Student's t GARCH (1,1)	2.6767	0.009817*	2.7224	0.003295*	2.2863	0.011220
Skewed Student's t GARCH (1,1)	2.4195	0.012017	2.2822	0.011341	2.9503	0.001624*
Normal EGARCH (1,1)	1.9534	0.025525	3.7374	0.000094*	5.3781	4.69E-08*
Student's t EGARCH (1,1)	1.4983	0.067175	3.7529	0.000093*	7.7801	8.99E-15*
Skewed Student's t EGARCH (1,1)	1.4592	0.072446	3.8885	0.000054*	7.8593	4.97E-15*
Normal GJR-GARCH (1,1)	2.2098	0.013661	3.4254	0.000312*	2.7161	0.003361*
Student's t GJR-GARCH (1,1)	1.5683	0.058559	2.9747	0.001501*	3.6485	0.000147*
Skewed Student's t GJR-GARCH (1,1)	1.4026	0.080523	3.0699	0.001099*	3.6576	0.000132*
Model	Period 4		Period 5		Period 6	
	Test statistic	P-value	Test statistic	P-value	Test statistic	P-value
Normal GARCH (1,1)	2.9716	0.001516*	5.1006	2.02E-07*	2.9249	0.001761*
Student's t GARCH (1,1)	2.9823	0.001464*	4.1741	0.000016*	3.5042	0.000239*
Skewed Student's t GARCH (1,1)	2.9883	0.001436*	4.1763	0.000016*	3.4895	0.000252*
Normal EGARCH (1,1)	2.8469	0.002253*	5.4736	2.77E-08*	2.2613	0.011977
Student's t EGARCH (1,1)	2.5737	0.005106*	5.0936	2.09E-07*	4.7987	9.16E-07*
Skewed Student's t EGARCH (1,1)	2.4634	0.006963*	5.1527	1.54E-07*	5.4796	2.68E-08*
Normal GJR-GARCH (1,1)	2.9632	0.001562*	5.5041	2.34E-08*	3.1871	0.000745*
Student's t GJR-GARCH (1,1)	3.1437	0.000867*	5.4798	2.68E-08*	5.0355	2.81E-07*
Skewed Student's t GJR-GARCH (1,1)	3.2937	0.000510*	5.5261	2.08E-08*	5.4944	2.47E-08*

* denote statistical significance at 1% level

Table 10 – Diebold-Mariano test statistic and p-value for high frequency data: Bitcoin

Model	Period 1		Period 2		Period 3	
	Test statistic	P-value	Test statistic	P-value	Test statistic	P-value
Normal GARCH (1,1)	2.4904	0.006455*	5.1662	1.43E-07*	2.2184	0.013390
Student's t GARCH (1,1)	3.1436	0.000859*	5.2612	8.70E-08*	2.4427	0.007375*
Skewed Student's t GARCH (1,1)	3.1527	0.000831*	5.2353	9.97E-08*	2.4426	0.007376*
Normal EGARCH (1,1)	2.4252	0.007731*	5.6387	1.11E-08*	2.8385	0.002313*
Student's t EGARCH (1,1)	3.6946	0.000116*	4.7779	0.000001*	2.5150	0.006032*
Skewed Student's t EGARCH (1,1)	3.7096	0.000114*	4.7754	0.00000*1	2.5169	0.000624*
Normal GJR-GARCH (1,1)	2.4415	0.007392*	2.3807	0.008736*	2.3923	0.008468*
Student's t GJR-GARCH (1,1)	3.8997	0.000051*	5.1359	1.69E-07*	2.8706	0.002094*
Skewed Student's t GJR-GARCH (1,1)	3.9118	0.000049*	5.2056	1.17E-07*	2.8692	0.002881*
Model	Period 4		Period 5		Period 6	
	Test statistic	P-value	Test statistic	P-value	Test statistic	P-value
Normal GARCH (1,1)	2.0124	0.022220	2.9990	0.001387*	4.2444	0.000012*
Student's t GARCH (1,1)	1.6976	0.044941	2.3534	0.009394*	4.7414	0.000001*
Skewed Student's t GARCH (1,1)	1.6328	0.051418	2.3480	0.009531*	4.7334	0.000001*
Normal EGARCH (1,1)	2.9144	0.001825*	1.7739	0.038192	4.1469	0.000018*
Student's t EGARCH (1,1)	2.5334	0.005725*	3.2161	0.000637*	6.0355	1.10E-09*
Skewed Student's t EGARCH (1,1)	2.5332	0.005734*	2.8446	0.002271*	6.0151	1.25E-09*
Normal GJR-GARCH (1,1)	3.2028	0.000716*	2.9896	0.001436*	4.3739	0.000006*
Student's t GJR-GARCH (1,1)	2.7264	0.003250*	2.2146	0.002027*	4.8830	6.05E-07*
Skewed Student's t GJR-GARCH (1,1)	2.3974	0.008356*	2.5355	0.005691*	5.0172	3.09E-07*

* denote statistical significance at 1% level

Table 11 – Diebold-Mariano test statistic and p-value for high frequency data: Ethereum

Model	Period 1		Period 2		Period 3	
	Test statistic	P-value	Test statistic	P-value	Test statistic	P-value
Normal GARCH (1,1)	2.7245	0.003273*	4.2093	0.000014*	2.2012	0.013980
Student's t GARCH (1,1)	2.3434	0.009646*	3.9008	0.000051*	3.2967	0.000506*
Skewed Student's t GARCH (1,1)	2.3229	0.010190	4.5353	0.000003*	3.2975	0.000505*
Normal EGARCH (1,1)	2.0813	0.018825	4.0735	0.000025*	4.2449	0.000012*
Student's t EGARCH (1,1)	2.0574	0.019944	5.2817	7.81E-08*	4.8814	6.13E-07*
Skewed Student's t EGARCH (1,1)	2.0735	0.019238	5.3128	6.61E-08*	5.3664	4.99E-08*
Normal GJR-GARCH (1,1)	2.9399	0.001677*	4.1774	0.000016*	1.7125	0.043657
Student's t GJR-GARCH (1,1)	2.3527	0.009408*	5.3468	5.51E-08*	4.8508	7.13E-07*
Skewed Student's t GJR-GARCH (1,1)	2.3424	0.009671*	5.3854	4.48E-08*	4.9205	5.04E-07*
Model	Period 4		Period 5		Period 6	
	Test statistic	P-value	Test statistic	P-value	Test statistic	P-value
Normal GARCH (1,1)	3.0978	0.001001*	4.1358	0.000019*	1.8299	0.033780
Student's t GARCH (1,1)	2.4952	0.006373*	4.0277	0.000030*	2.3599	0.009234*
Skewed Student's t GARCH (1,1)	2.4955	0.006368*	3.9982	0.000034*	2.5752	0.005079*
Normal EGARCH (1,1)	3.3524	0.001825*	1.7739	0.038198	4.3126	0.000008*
Student's t EGARCH (1,1)	6.0762	8.65E-10*	4.9447	4.45E-07*	5.4134	3.85E-08*
Skewed Student's t EGARCH (1,1)	6.1052	7.26E-10*	4.9337	4.70E-07*	4.6512	0.000002*
Normal GJR-GARCH (1,1)	3.2028	0.000701*	2.9896	0.001438*	9.0311	2.21E-16*
Student's t GJR-GARCH (1,1)	2.4603	0.007026*	3.2242	0.000652*	5.4175	3.76E-08*
Skewed Student's t GJR-GARCH (1,1)	2.4598	0.007039*	3.2027	0.000702*	6.1086	7.12E-10*

* denote statistical significance at 1% level

Table 12 – Diebold-Mariano test statistic and p-value for high frequency data: Dashcoin

Model	Period 1		Period 2		Period 3	
	Test statistic	P-value	Test statistic	P-value	Test statistic	P-value
Normal GARCH (1,1)	3.1210	0.000926*	3.5607	0.000193*	4.1963	0.000015*
Student's t GARCH (1,1)	3.0174	0.001306*	3.7950	0.000078*	4.3448	0.000008*
Skewed Student's t GARCH (1,1)	3.3358	0.000441*	3.4473	0.000295*	5.9567	1.78E-09*
Normal EGARCH (1,1)	2.7713	0.002846*	2.7242	0.003278*	5.7828	4.93E-09*
Student's t EGARCH (1,1)	2.4172	0.007903*	4.7246	0.000001*	5.7665	5.39E-09*
Skewed Student's t EGARCH (1,1)	2.4548	0.007126*	4.7798	0.000001*	5.9810	1.54E-09*
Normal GJR-GARCH (1,1)	4.3344	0.000008*	3.4908	0.000251*	5.8684	2.99E-09*
Student's t GJR-GARCH (1,1)	3.6091	0.000161*	4.9291	4.81E-07*	5.6071	1.33E-08*
Skewed Student's t GJR-GARCH (1,1)	3.4175	0.000328*	4.0377	0.000029*	5.4033	4.09E-08*
Model	Period 4		Period 5		Period 6	
	Test statistic	P-value	Test statistic	P-value	Test statistic	P-value
Normal GARCH (1,1)	2.6110	0.004586*	5.3216	6.31E-08*	5.8903	2.61E-09*
Student's t GARCH (1,1)	2.7509	0.003024*	5.2231	1.06E-07*	4.9646	4.03E-07*
Skewed Student's t GARCH (1,1)	2.4682	0.006871*	6.3077	2.10E-10*	5.4034	4.06E-08*
Normal EGARCH (1,1)	1.9437	0.026193	1.6794	0.046682	3.8213	0.000073*
Student's t EGARCH (1,1)	3.9026	0.000051*	2.5272	0.005817*	6.7386	1.33E-11*
Skewed Student's t EGARCH (1,1)	3.8718	0.000057*	2.5005	0.006282*	6.4515	8.51E-11*
Normal GJR-GARCH (1,1)	5.4734	2.71E-08*	2.5198	0.005956*	7.2867	3.16E-13*
Student's t GJR-GARCH (1,1)	1.3995	0.080991	1.8606	0.031547	3.7447	0.000095*
Skewed Student's t GJR-GARCH (1,1)	3.7595	0.000089*	2.0243	0.021617	1.9203	0.027554

* denote statistical significance at 1% level

Table 13 – Diebold-Mariano test statistic and p-value for high frequency data: Euro

Model	Period 1		Period 2		Period 3	
	Test statistic	P-value	Test statistic	P-value	Test statistic	P-value
Normal GARCH (1,1)	4.3898	0.000006*	3.2182	0.000665*	4.7147	0.000001*
Student's t GARCH (1,1)	6.0130	1.25E-09*	3.6054	0.000163*	5.3017	7.06E-08*
Skewed Student's t GARCH (1,1)	5.4384	3.34E-08*	3.6730	0.000126*	5.7849	4.85E-09*
Normal EGARCH (1,1)	3.6958	0.000115*	3.8564	0.000061*	2.5158	0.006016*
Student's t EGARCH (1,1)	4.1089	0.000021*	2.5209	0.005935*	4.4156	0.000005*
Skewed Student's t EGARCH (1,1)	4.0913	0.000023*	2.5102	0.006119*	4.3022	0.000009*
Normal GJR-GARCH (1,1)	5.9356	1.97E-09*	1.6499	0.049632	4.8843	6.04E-07*
Student's t GJR-GARCH (1,1)	3.4149	0.000331*	2.4856	0.006546*	5.8732	2.91E-09*
Skewed Student's t GJR-GARCH (1,1)	7.6605	2.06E-14*	2.2931	0.011028	5.6831	8.67E-09*
Model	Period 4		Period 5		Period 6	
	Test statistic	P-value	Test statistic	P-value	Test statistic	P-value
Normal GARCH (1,1)	3.4275	0.000316*	5.5846	1.50E-08*	2.2302	0.012970
Student's t GARCH (1,1)	4.0127	0.000032*	5.2628	8.63E-08*	2.2913	0.011110
Skewed Student's t GARCH (1,1)	3.7121	0.000108*	5.7619	5.49E-09*	1.9615	0.025043
Normal EGARCH (1,1)	2.7495	0.002953*	4.7792	0.000001*	2.3605	0.009221*
Student's t EGARCH (1,1)	8.6976	2.21E-16*	6.9767	2.71E-12*	1.2961	0.097612
Skewed Student's t EGARCH (1,1)	8.7316	2.21E-16*	6.8923	4.79E-12*	1.3113	0.095034
Normal GJR-GARCH (1,1)	4.4363	0.000005*	5.2254	1.05E-07*	1.1832	0.118517
Student's t GJR-GARCH (1,1)	8.3219	2.21E-16*	4.9438	4.47E-07*	2.1657	0.012863
Skewed Student's t GJR-GARCH (1,1)	6.2295	3.45E-10*	4.7118	0.000001*	2.6445	0.004153*

* denote statistical significance at 1% level

Table 14 – Diebold-Mariano test statistic and p-value for high frequency data: British Pound

Model	Period 1		Period 2		Period 3	
	Test statistic	P-value	Test statistic	P-value	Test statistic	P-value
Normal GARCH (1,1)	4.9330	4.69E-07*	3.5532	0.000199*	2.7268	0.003254*
Student's t GARCH (1,1)	5.6116	1.28E-08*	3.0526	0.001163*	2.5908	0.004857*
Skewed Student's t GARCH (1,1)	4.6434	1.93E-06*	3.0732	0.001087*	2.4158	0.000794*
Normal EGARCH (1,1)	6.2383	3.17E-10*	3.3271	0.000454*	2.0245	0.021597
Student's t EGARCH (1,1)	7.2093	5.29E-13*	3.8657	0.000059*	3.1253	0.000914*
Skewed Student's t EGARCH (1,1)	7.2378	4.33E-13*	3.8422	0.000065*	3.2885	0.000521*
Normal GJR-GARCH (1,1)	8.5345	2.21E-16*	3.3737	0.003853*	3.3127	0.000479*
Student's t GJR-GARCH (1,1)	3.3764	0.000098*	5.2041	1.18E-07*	2.5059	0.006187*
Skewed Student's t GJR-GARCH (1,1)	6.0879	7.94E-10*	1.0219	0.153517	1.6333	0.051369
Model	Period 4		Period 5		Period 6	
	Test statistic	P-value	Test statistic	P-value	Test statistic	P-value
Normal GARCH (1,1)	4.7986	9.17E-07*	2.9258	0.001756*	3.5971	0.000168*
Student's t GARCH (1,1)	4.8939	5.73E-07*	2.7229	0.003290*	4.2148	0.000014*
Skewed Student's t GARCH (1,1)	4.9584	4.15E-07*	3.0272	0.001265*	4.4181	0.000006*
Normal EGARCH (1,1)	4.9784	3.76E-07*	1.6405	0.050607	3.9924	0.000035*
Student's t EGARCH (1,1)	5.0017	3.34E-07*	1.6904	0.045638	4.3206	0.000008*
Skewed Student's t EGARCH (1,1)	5.0091	3.22E-07*	1.4656	0.071532	4.2961	0.000016*
Normal GJR-GARCH (1,1)	6.6291	2.72E-11*	3.0451	0.001193*	3.9831	0.000036*
Student's t GJR-GARCH (1,1)	4.5513	0.000003*	1.9769	0.024166	3.3566	0.000409*
Skewed Student's t GJR-GARCH (1,1)	3.2319	0.000634*	1.1124	0.133183	2.9392	0.001685*

* denote statistical significance at 1% level

Table 15 – Diebold-Mariano test statistic and p-value for high frequency data: Japanese Yen

3.5 Results and discussion

The results were widely favorable towards SVR-GARCH model. As shown in tables **2** to **8**, both RMSE and MAE of SVR-GARCH were lower than Normal, Student's t and Skewed Student's t distributions GARCHs, EGARCHs and GJR-GARCHs for all exchange rates and cryptocurrencies, in both low frequency and high frequency datasets. Similarly, the Diebold-Mariano test and Hansen's superior model sets also present strong evidences that the SVR models significantly outperforms the traditional GARCH models.

As shown in tables **9** to **15**, at the usual 95% confidence level only 19 out of 420 models failed to reject the null hypothesis, indicating that SVR models performed much better than the other models during all data range in both time frequencies for all assets – even at the 99% confidence level, SVR-GARCH showed predictive superiority in 319 out of 420 models: 42 out of 60 models for low frequency and for 277 out of 360 models for high frequency data.

As for Hansen, Lunde and Nason (2011)'s model confidence set procedure, the SVR-GARCH model was in the set of superior models for all assets and all time periods – in fact, almost all SSMs (35 out of 42 sets) were composed only by the SVR-GARCH model, which provide further evidences the better performance of the machine learning approach. Overall, the SSMs generated by the MCS tests contained a small number of models due to the relatively small size of the initial set $\mathcal{M}^0$ (10 models). In addition, SVR-GARCH's RMSE and MAE were in general significantly lower than all other benchmark models, making it "superior" to the rest by a wide margin in many cases.

Overall, the forecasting errors of both GARCH and SVR models were higher for the cryptocurrencies than for the traditional currencies in both daily and hourly frequencies, an expected outcome given the big difference in the volatility levels for those kinds of asset.

For the low frequency dataset, the error metrics were generally higher than the high frequency one. This result is consistent with the findings in the literature: as seen in Xie and Li (2010), the RMSE for the volatility forecasting tend to decrease as the frequency increases, a behavior that was also identified for virtual currencies. As shown by the Diebold-Mariano test p-values and Hansen's superior model sets, the SVR models seemed to outperform GARCH models less emphatically for the cryptocurrencies than for the exchange rates for the daily volatility, whilst for the hourly volatility the superiority evidences were stronger for cryptocurrencies and milder for traditional currencies. This suggests a higher intraday volatility fluctuation in exchange rates than for cryptocurrencies, which can be associated to the huge liquidity of the foreign exchange market in comparison to the incipient acceptability of bitcoin, ethereum and dash.

The good predictive performance of the SVR based models can be linked to the

nonlinearities that the Kernel function bring forth, inducing an infinite-dimensional feature space with a small number of parameters and incorporating nonlinear interactions that traditional linear models fail to capture. For assets with much higher volatility levels, like cryptocurrencies, the evidences of SVR's better predictive power are still very strong, suggesting the robustness of machine learning techniques in forecasting financial time series.

Analyzing the error metrics, the Diebold-Mariano test p-values and the composition of the set of superior models, the GJR-GARCH models seemed to perform slightly better than GARCH and EGARCH. Concerning the conditional distribution of ϵ_t, the Normal, Student's t and skewed Student's t distributions yielded overall similar results.

The results reveal that SVR models presented a significantly lower value for both RMSE and MAE error metrics in comparison to all nine GARCH models. The results are similar to the findings of Hsu et al. (2016), in which the authors conclude, based on various experiments, that machine learning techniques demonstrate superior predictive power than traditional econometric models. The results of this paper present evidences of such superiority not only for the exchange rates' volatility estimation, but also for the cryptocurrencies, a segment not explored by Hsu et al. (2016) and still not frequently studied in the finance literature.

Nonetheless, it was possible to see that cryptocurrencies have higher overall volatility than real world currencies. This result bring forth the discussion of whether cryptocurrencies can be treated as traditional currencies, specially concerning the debate regarding their fundamental or speculative nature, as discussed by Dowd (2014). The absence of a central monetary authority, while being the main feature of cryptocurrencies, is also one of the main sources of criticism, since a big share of their value is based on their notoriety and circulation on web, which creates a speculative profile for this kind of asset and a potential bubble. Authors like Baek and Elbeck (2015) expected that the increase in bitcoin's usage would make its volatility drop and exhibit a more investment-like behavior rather than a speculative tool. However, this did not happen, since bitcoin's market cap has been increasing all along, but its volatility is also going up, as seen in figure **2**. Thus, the higher volatility levels showed by cryptocurrencies suggests a more cautious look over cryptocurrencies, and presents a possible evidence that they cannot yet be considered as "trusted currencies", mainly because their lack of maturity upon the store of value function.

3.6 Conclusion and remarks

This paper evaluated SVR-GARCH's predictive performance of daily and hourly volatiilty of three cryptocurrencies and three exchange rate pairs. The GARCH model was

combined with machine learning approach, such that the mean and volatility equations were estimated using Support Vector Regression. Furthermore, we compared the models' predictive ability with Diebold-Mariano test and Hansen's Model Confidence Set. The results show that SVR-GARCH models managed to outperform all nine GARCH benchmarks – GARCHs, EGARCHs and GJR-GARCHs with Normal, Student's t and Skewed Student's t distributions – as seen by the value of error metrics RMSE and MAE, the Diebold-Mariano test p-values and the composition of the set of superior models.

The findings of this paper have the potential to aid scholars and market practitioners with an overview of the cryptocurrencies market features, discussing similarities and differences of their volatility patterns in comparison to real world currencies, presenting the extents in which the incorporation of a machine learning based technique yields better forecasting power for volatility over the GARCH benchmarks. The outcome of this research is a tool capable of estimating the risk for the cryptocurrencies in the future and can be used as a risk management for portfolios, as proposed by Dyhrberg (2016a). Furthermore, a more accurate model for volatility forecast in cryptocurrencies can be of interest for companies that accept them, as well as potential investors and traders of this market segment. More precise volatility predictions for cryptocurrencies can represent a measure of short-term risk to better evaluate the attractiveness of this kind of assets over alternative risky investments, potentially leading to better portfolio allocations, and guidance to investment decisions or corporate strategies.

Future researches are encouraged to replicate this study for other financial assets' volatility estimation, as well as to consider other distributions for the GARCH models' error term - such as Generalized Pareto Distribution (MCNEIL; FREY, 2000) - and other well known models for volatility estimation, such as TGARCH (ZAKOIAN, 1994) and APARCH (DING; GRANGER; ENGLE, 1993). Also, the inclusion of SVR estimation to other volatility models apart from GARCH (1,1) can further contribute to better volatility predictions and may be an attractive and relevant issue in future developments in the finance literature.

Bearing in mind the huge popularity and prominence of Machine Learning methods many scientific fields, finance included, testing for different extensions of SVR – like Chang and Lin (2002)'s ν-SVR – is also quite desirable. The use of different Kernel functions or mixture models, as seen in Bezerra and Albuquerque (2017), can also be incorporated to the SVR models. Finally, replications with different time periods and frequencies (*e.g.:* even higher frequency data, by minutes or even seconds) can contribute further for this research agenda.

Appendix A. Predictive Accuracy Test

Diebold and Mariano (1995)'s predictive accuracy test compares the loss differential between the forecasting errors of two sets relative to the observed values. We used SVR-GARCH model as benchmark, so we defined d_t as the excess error of SVR-GARCH model over the other GARCH models:

$$d_t = [g(e_{SVR,t}) - g(e_{GARCH,t})] \tag{3.15}$$

where $e_{i,t} = \tilde{h}_t - \widehat{h}_{i,t}$ is the forecast error of the i-th model at time t and $g(.)$ is a loss function, which we defined as the squared error $g(e_{i,t}) = e_{i,t}^2$.

The Diebold-Mariano test evaluates the null hypothesis

$$H_0 : \mathbb{E}[d_t] \geq 0, \quad \forall\, t = 1, 2, ..., T \tag{3.16}$$

where T is the number of time periods in the test sets (thus, the number of forecasts generated). The null hypothesis states that SVR-GARCH models have equal or worse accuracy than GARCH models, while its rejection provides evidence of superiority over them.

Appendix B. Model Confidence Set

Hansen, Lunde and Nason (2011)'s Model Confidence Set (MCS) provides, at a given significance level α, a subset of "superior models" from an initial set $\mathcal{M}^0$ containing all m tested models. The superior set models (SSM) is obtained by recursively removing the worst model in $\mathcal{M}^0$ evaluating the null hypothesis of equal predictive ability for the i-th model in $\mathcal{M}^0, i = 1, 2, ..., m$, given by:

$$H_0 : \quad \mathbb{E}\left[\sum_{j=1}^{m-1}\sum_{t=1}^{T} g(e_{i,t}) - g(e_{j,t})\right] = 0, \quad i = 1, 2, ..., m \tag{3.17}$$

with T, $g(.)$ and $e_{i,t}$ as previously defined for the Diebold-Mariano test.

The MCS procedure basically tests whether the models in the initial set $\mathcal{M}^0$ have equal predictive power (null hypothesis); a block bootstrap procedure is used to compute the distribution under H_0. If H_0 is not rejected, then $\mathcal{M}^0$ is itself the superior set of models SSM. On the other hand, if H_0 is rejected, then at least one model "differs significantly" from the others in terms of predictive quality, so that one of the m models in $\mathcal{M}^0$ is chosen as the worst model and eliminated from $\mathcal{M}^0$, filtering it into a subset $\mathcal{M}^*$ containing "better models". The elimination rule removes the model with the worst

relative performance in comparison to the average across all other models, measured by the test statistic:

$$t_i = \frac{\frac{1}{m-1}\sum_{j=1}^{m-1}\sum_{t=1}^{T} g(e_{i,t}) - g(e_{j,t})}{\sqrt{\widehat{\mathbb{V}}\left(\frac{1}{m-1}\sum_{j=1}^{m-1}\sum_{t=1}^{T} g(e_{i,t}) - g(e_{j,t})\right)}} \tag{3.18}$$

The routine is applied recursively, eliminating one model at a time. Each time the equal predictive ability null hypothesis is rejected, $\mathcal{M}^*$ is updated, eliminating the worst models. When H_0 ceases to be rejected, the current $\mathcal{M}^*$ is the SSM containing only the "superior models" at the significance level α.

4 Between Nonlinearities, Complexity, and Noises: An Application on Portfolio Selection Using Kernel Principal Component Analysis

Abstract

This paper discusses the effects of introducing nonlinear interactions and noise-filtering to the covariance matrix used in Markowitz's portfolio allocation model, evaluating the technique's performances for daily data from seven financial markets between January 2000 and August 2018. We estimated the covariance matrix by applying Kernel functions, and applied filtering following the theoretical distribution of the eigenvalues based on the Random Matrix Theory. The results were compared with the traditional linear Pearson estimator and robust estimation methods for covariance matrices. The results showed that noise-filtering yielded portfolios with significantly larger risk-adjusted profitability than its non-filtered counterpart for almost half of the tested cases. Moreover, we analyzed the improvements and setbacks of the nonlinear approaches over linear ones, discussing in which circumstances the additional complexity of nonlinear features seemed to predominantly add more noise or predictive performance.[1]

4.1 Introduction

Finance can be defined as the research field that studies the management of value—for an arbitrary investor that operates inside the financial market, the value of the assets that he/she chose can be measured in terms of how profitable or risky they are. While individuals tend to pursue potentially larger return rates, often the most profitable options bring along higher levels of uncertainty as well, so that the risk–return relationship induces a trade-off over the preferences of the economic agents, making them seek a combination of assets that offer maximum profitability, as well as minimum risk—an efficient allocation of the resources that generate the most payoff/reward/value.

As pointed out in Miller (1999), one of the main milestones in the history of finance was the mean-variance model of Nobel Prize laureate Harry Markowitz, a work regarded as the genesis of the so-called "Modern Portfolio Theory", in which the optimal portfolio

[1] Published in *Entropy*, v. 21, n. 4, p. 376, 2019

choice was presented as the solution of a simple, constrained optimization problem. Furthermore, Markowitz (1952)'s model shows the circumstances in which the levels of risk can be diminished through diversification, as well as the limits of this artifice, represented by a risk that investors can do nothing about and therefore must take when investing in the financial market.

While the relevance of Markowitz (1952)'s work is unanimously praised, the best way to estimate its inputs—a vector of expected returns and a covariance matrix—is far from reaching a consensus. While the standard estimators are easy to obtain, recent works like Pavlidis, Paya and Peel (2015) and Hsu et al. (2016) argue in favor of the introduction of nonlinear features to boost the predictive power for financial variables over traditional parametric econometric methods, and in which existing novel approaches, such as machine-learning methods, can contribute to better forecasting performances. Additionally, many studies globally have found empirical evidence from real-world financial data that the underlying patterns of financial covariance matrices seem to follow some stylized facts regarding the big proportion of "noise" in comparison to actually useful information, implying that the complexity of the portfolio choice problem could be largely reduced, possibly leading to more parsimonious models that provide better forecasts.

This paper focused on those questions, investigating whether the use of a nonlinear and nonparametric covariance matrix or the application of noise-filtering techniques can indeed help a financial investor to build better portfolios in terms of cumulative return and risk-adjusted measures, namely Sharpe and Sortino ratios. Moreover, we analyzed various robust methods for estimating the covariance matrix, and whether nonlinearities and noise-filtering managed to bring improvements to the portfolios' performance, which can be useful to the construction of portfolio-building strategies for financial investors. We tested different markets and compared the results, and discussed to which extent the portfolio allocation was done better using Kernel functions and "clean" covariance matrices.

The paper is structured as follows: Section 4.2 presents the foundations of risk diversification via portfolios, discussing the issues regarding high dimensionality in financial data, motivating the use of high-frequency data, as well as nonlinear predictors, regularization techniques, and the Random Matrix Theory. Section 4.3 describes the Markowitz (1952) portfolio selection model, robust estimators for the covariance matrix, and the Principal Component Analysis for both linear and Kernel covariance matrices. Section 4.4 provides details on the empirical analysis and describes the collected data and chosen time periods, as well as the performance metrics and statistical tests for the evaluation of the portfolio allocations. Section 4.5 presents the performance of the obtained portfolios and discusses their implication in view of the financial theory. Finally, Section 4.6 presents the paper's conclusions, potential limitations to the proposed methods, and

recommendations for future developments.

4.2 Theoretical Background

4.2.1 Portfolio Selection and Risk Management

In financial contexts, "risk" refers to the likelihood of an investment yielding a different return from the expected one (DAMODARAN, 2012); thus, in a broad sense, risk does not necessarily only have regard to unfavorable outcomes (downside risk), but rather includes upside risk as well. Any flotation from the expected value of the return of a financial asset is viewed as a source of uncertainty, or "volatility", as it is more often called in finance.

A rational investor would seek to optimize his interests at all times, which can be expressed in terms of maximization of his expected return and minimization of his risk. Given that future returns are a random variable, there are many possible measures for its volatility; however, the most common measure for risk is the variance operator (second moment), as used in Markowitz (1952)'s Modern Portfolio Theory seminal work, while expected return is measured by the first moment. This is equivalent to assuming that all financial agents follow a mean-variance preference, which is grounded in the microeconomic theory and has implications in the derivation of many important models in finance and asset pricing, such as the CAPM model (SHARPE, 1964; LINTNER, 1965; MOSSIN, 1966), for instance.

The assumption of rationality implies that an "efficient" portfolio allocation is a choice of weights $\boldsymbol{w}$ in regard to how much assets you should buy which are available in the market, such that the investor cannot increase his expected return without taking more risk—or, alternatively, how you can decrease his portfolio volatility without taking a lower level of expected return. The curve of the possible efficient portfolio allocations in the risk versus the expected return graph is known as an "efficient frontier". As shown in Markowitz (1952), in order to achieve an efficient portfolio, the investor should diversify his/her choices, picking the assets with the minimal association (measured by covariances), such that the joint risks of the picked assets tend to cancel each other.

Therefore, for a set of assets with identical values for expected return μ and variance σ^2, choosing a convex combination of many of them will yield a portfolio with a volatility value smaller than σ^2, unless all chosen assets have perfect correlation. Such effects of diversification can be seen statistically from the variance of the sum of p random variables: $\mathbb{V}[w_1X_1 + w_2X_2 + ... + w_pX_p] = \sum_{i=1}^{p}\sum_{j=1}^{p} w_i w_j cov(X_i, X_j)$; since $\sum_{i=1}^{p} w_i = 1$ (negative-valued weights represent a short selling), the volatility of a generic portfolio $w_1x_1 + w_2x_2 + ... + w_px_p$ with same-risk assets will always diminish with diversification.

The component of risk which can be diversified, corresponding to the joint volatility between the chosen assets, is known as "idiosyncratic risk", while the non-diversifiable component of risk, which represents the uncertainties associated to the financial market itself, is known as "systematic risk" or "market risk". The idiosyncratic risk is specific to a company, industry, market, economy, or country, meaning it can be eliminated by simply investing in different assets (diversification) that will not all be affected in the same way by market events. On the other hand, the market risk is associated with factors that affect all assets' companies, such as macroeconomic indicators and political scenarios; thus not being specific to a particular company or industry and which cannot be eliminated or reduced through diversification.

Although there are many influential portfolio selection models that arose after Markowitz's classic work, such as the Treynor-Black model (TREYNOR; BLACK, 1973), the Black-Litterman model (BLACK; LITTERMAN, 1992), as well as advances in the so-called "Post-Modern Portfolio Theory" (ROM; FERGUSON, 1994; GALLOPPO et al., 2010) and machine-learning techniques (ZHANG; ZHANG; XIAO, 2009; HUANG, 2012; MARCELINO; HENRIQUE; ALBUQUERQUE, 2015), Markowitz (1952) remains as one of the most influential works in finance and is still widely used as a benchmark for alternative portfolio selection models, due to its mathematical simplicity (uses only a vector of expected returns and a covariance matrix as inputs) and easiness of interpretation. Therefore, we used this model as a baseline to explore the potential improvements that arise with the introduction of nonlinear interactions and covariance matrix filtering through the Random Matrix Theory.

4.2.2 Nonlinearities and Machine Learning in Financial Applications

Buonocore et al. (2016) presents two key elements that define the complexity of financial time-series: the multi-scaling property, which refers to the dynamics of the series over time; and the structure of cross-dependence between time-series, which are reflexes of the interactions among the various financial assets and economic agents. In a financial context, one can view those two complexity elements as systematic risk and idiosyncratic risk, respectively, precisely being the two sources of risk that drive the whole motivation for risk diversification via portfolio allocation, as discussed by the Modern Portfolio Theory.

It is well-known that systematic risk cannot be diversified. So, in terms of risk management and portfolio selection, the main issue is to pick assets with minimal idiosyncratic risk, which in turn, naturally, demands a good estimation for the cross-interaction between the assets available in the market, namely the covariance between them.

The non-stationarity of financial time-series is a stylized fact which is well-known by scholars and market practitioners, and this property has relevant implications in forecasting and identifying patterns in financial analysis. Specifically concerning portfolio

selection, the non-stationary behavior of stock prices can induce major drawbacks when using the standard linear Pearson correlation estimator in calculating the covariances matrix. Livan, Inoue and Scalas (2012) provides empirical evidence of the limitations of the traditional linear approach established in Markowitz (1952), pointing out that the linear estimator fails to accurately capture the market's dynamics over time, an issue that is not efficiently solved by simply using a longer historical series. The sensitivity of Markowitz (1952)'s model to its inputs is also discussed in Chen and Zhou (2018), which incorporates the third and fourth moments (skewness and kurtosis) as additional sources of uncertainty over the variance. Using multi-objective particle swarm optimization, robust efficient portfolios were obtained and shown to improve the expected return in comparison to the traditional mean-variance approach. The relative attractiveness of different robust efficient solutions to different market settings (bullish, steady, and bearish) was also discussed.

Concerning the Dynamical Behavior of Financial Systems, Bonanno, Valenti and Spagnolo (2006) proposed a generalization of the Heston model (HESTON, 1993), which is defined by two coupled stochastic differential equations (SDEs) representing the log of the price levels and the volatility of financial stocks, and provided a solution for option pricing that incorporated improvements over the classical Black-Scholes model (BLACK; SCHOLES, 1973) regarding financial stylized facts, such as the skewness of the returns and the excess kurtosis. The extension proposed by Bonanno, Valenti and Spagnolo (2006) was the introduction of a random walk with cubic nonlinearity to replace the log-price SDE of Heston's model. Furthermore, the authors analyzed the statistical properties of escape time as a measure of the stabilizing effect of the noise in the market dynamics. Applying this extended model, Spagnolo and Valenti (2008) tested for daily data of 1071 stocks traded at the New York Stock Exchange between 1987 and 1998, finding out that the nonlinear Heston model approximates the probability density distribution on escape times better than the basic geometric Brownian motion model and two well-known volatility models, namely GARCH (BOLLERSLEV, 1986) and the original Heston model (HESTON, 1993). In this way, the introduction of a nonlinear term allowed for a better understanding of a measure of market instability, capturing embedded relationships that linear estimators fail to consider. Similarly, linear estimators for covariance ignore potential associations in higher dimensionality interactions, such that even assets with zero covariance may actually have a very heavy dependence on nonlinear domains.

As discussed in Kühn and Neu (2008), the states of a market can be viewed as attractors resulting from the dynamics of nonlinear interactions between the financial variables, such that the introduction of nonlinearities also has potential implications for financial applications, such as risk management and derivatives pricing. For instance, Valenti, Fazio and Spagnolo (2018) pointed out that volatility is a monotonic indicator of financial risk, while many large oscillations in a financial market (both upwards and

downwards) are preceded by long periods of relatively small levels of volatility in the assets' returns (the so-called "volatility clustering"). In this sense, the authors proposed the mean first hitting time (defined as the average time until a stock return undergoes a large variation—positive or negative—for the first time) as an indicator of price stability. In contrast with volatility, this measure of stability displays nonmonotonic behavior that exhibits a pattern resembling the Noise Enhanced Stability (NES) phenomenon, observed in a broad class of systems (AGUDOV; DUBKOV; SPAGNOLO, 2003; DUBKOV; AGUDOV; SPAGNOLO, 2004; FIASCONARO; SPAGNOLO; BOCCALETTI, 2005). Therefore, using the conventional volatility as a measure of risk can lead to its underestimation, which in turn can lead to bad allocations of resources or bad financial managerial decisions.

In light of evidence that not all noisy information of the covariance matrix is due to their non-stationarity behavior (MARTINS, 2007), many machine-learning methods, such as the Support Vector Machines (GUPTA; MEHLAWAT; MITTAL, 2012), Gaussian processes (PARK et al., 2016), and deep learning (HEATON; POLSON; WITTE, 2017) methods have been discussed in the literature, showing that the introduction of nonlinearities can provide a better display of the complex cross-interactions between the variables and generate better predictions and strategies for the financial markets. Similarly, Almahdi and Yang (2017) proposed a portfolio trading algorithm using recurrent reinforcement learning, using the expected maximum drawdown as a downside risk measure and testing for different sets of transaction costs. The authors also proposed an adaptive rebalancing extension, reported to have a quicker reaction to transaction cost variations and which managed to outperform hedge fund benchmarks.

Paiva et al. (2018) proposed a fusion approach of a Support Vector Machine and the mean-variance optimization for portfolio selection, testing for data from the Brazilian market and analyzing the effects of brokerage and transactions costs. Petropoulos et al. (2017) applied five machine learning algorithms (Support Vector Machine, Random Forest, Deep Artificial Neural Networks, Bayesian Autoregressive Trees, and Naïve Bayes) to build a model for FOREX portfolio management, combining the aforementioned methods in a stacked generalization system. Testing for data from 2001 to 2015 of ten currency pairs, the authors reported the superiority of machine learning models in terms of out-of-sample profitability. Moreover, the paper discussed potential correlations between the individual machine learning models, providing insights concerning their combination to boost the overall predictive power. Chen et al. (2009) generalized the idea of diversifying for individual assets for investment and proposed a framework to construct portfolios of investment strategies instead. The authors used genetic algorithms to find the optimal allocation of capital into different strategies. For an overview of the applications of machine learning techniques in portfolio management contexts, see Pareek and Thakkar (2015).

Regarding portfolio selection, Chicheportiche and Bouchaud (2015) developed a nested factor multivariate model to model the nonlinear interactions in stock returns, as well as the well-known stylized facts and empirically detected copula structures. Testing for the S&P 500 index for three time periods (before, during, and after the financial crisis), the paper showed that the optimal portfolio constructed by the developed model showed a significantly lower out-of-sample risk than the one built using linear Principal Component Analysis, whilst the in-sample risk is practically the same; thus being positive evidence towards the introduction of nonlinearities in portfolio selection and asset allocation models. Montenegro and Albuquerque (2017) applied a local Gaussian correlation to model the nonlinear dependence structure of the dynamic relationship between the assets. Using a subset of companies from the S&P 500 Index between 1992 and 2015, the portfolio generated by the nonlinear approach managed to outperform the Markowitz (1952) model in more than 60% of the validation bootstrap samples. In regard to the effects of dimensionality reduction on the performance of portfolios generated from mean-variance optimization, Tayalı and Tolun (2018) applied Non-negative Matrix Factorization (NMF) and Non-negative Principal Components Analysis (NPCA) for data from three indexes of the Istanbul Stock Market. Optimal portfolios were constructed based on Markowitz (1952)'s mean-variance model. Performing backtesting for 300 tangency portfolios (maximum Sharpe Ratio), the authors showed that the portfolios' efficiency was improved in both NMF and NPCA approaches over the unreduced covariance matrix.

Musmeci, Aste and Matteo (2016) incorporated a metric of persistence in the correlation structure between financial assets, and argued that such persistence can be useful for the anticipation of market volatility variations and that they could quickly adapt to them. Testing for daily prices of US and UK stocks between 1997 and 2013, the correlation structure persistence model yielded better forecasts than predictors based exclusively on past volatility. Moreover, the paper discusses the effect of the "curse of dimensionality" that arises in financial data when a large number of assets is considered, an issue that traditional econometric methods often fail to deal with. In this regard, Hsu et al. (2016) argues in favor of the use of nonparametric approaches and machine learning methods in traditional financial economics problems, given their better empirical predictive power, as well as providing a broader view of well-established research topics in the finance agenda beyond classic econometrics.

4.2.3 Regularization, Noise Filtering, and Random Matrix Theory

A major setback in introducing nonlinearities is keeping them under control, as they tend to significantly boost the model's complexity, both in terms of theoretical implications and computational power needed to actually perform the calculations. Nonlinear interactions, besides often being difficult to interpret and apart from a potentially better

explanatory power, may bring alongside them a large amount of noisy information, such as an increase in complexity that is not compensated by better forecasts or theoretical insights, but instead which "pollutes" the model by filling it with potentially useless data.

Bearing in mind this setback, the presence of regularization is essential to cope with the complexity levels that come along with high dimensionality and nonlinear interactions, especially in financial applications in which the data-generating processes tend to be highly chaotic. While it is important to introduce new sources of potentially useful information by boosting the model's complexity, being able to filter that information, discard the noises, and maintain only the "good" information is a big and relevant challenge. Studies like Massara, Matteo and Aste (2016) discuss the importance of scalability and information filtering in light of the advent of the "Big Data Era", in which the boost of data availability and abundance led to the need to efficiently use those data and filter out the redundant ones.

Barfuss et al. (2016) emphasized the need for parsimonious models by using information filtering networks, and building sparse-structure models that showed similar predictive performances but much smaller computational processing time in comparison to a state-of-the-art sparse graphical model baseline. Similarly, Torun, Akansu and Avellaneda (2011) discussed the eigenfiltering of measurement noise for hedged portfolios, showing that empirically estimated financial correlation matrices contain high levels of intrinsic noise, and proposed several methods for filtering it in risk engineering applications.

In financial contexts, Ban, Karoui and Lim (2016) discussed the effects of performance-based regularization in portfolio optimization for mean-variance and mean-conditional Value-at-Risk problems, showing evidence for its superiority towards traditional optimization and regularization methods in terms of diminishing the estimation error and shrinking the model's overall complexity.

Concerning the effects of high dimensionality in finance, Kozak, Nagel and Santosh (2017) tested many well-established asset pricing factor models (including CAPM and the Fama-French five-factor model) introducing nonlinear interactions between 50 anomaly characteristics and 80 financial ratios up to the third power (i.e., all cross-interactions between the features of first, second, and third degrees were included as predictors, totaling to models with 1375 and 3400 candidate factors, respectively). In order to shrink the complexity of the model's high dimensionality, the authors applied dimensionality reduction and regularization techniques considering ℓ_1 and ℓ_2 penalties to increase the model's sparsity. The results showed that a very small number of principal components were able to capture almost all of the out-of-sample explanatory powers, resulting in a much more parsimonious and easy-to-interpret model; moreover, the introduction of an additional regularized principal component was shown to not hinder the model's sparsity, but also to not improve predictive performance either.

Depending on the "noisiness" of the data, the estimation of the covariances can be severely hindered, potentially leading to bad portfolio allocation decisions—if the covariances are overestimated, the investor could give up less risky asset combinations, or accept a lesser expected profitability; if the covariances are underestimated, the investor would be bearing a higher risk than the level he was willing to accept, and his portfolio choice could be non-optimal in terms of risk and return. Livan, Inoue and Scalas (2012) discussed the impacts of measurement noises on correlation estimates and the desirability of filtering and regularization techniques to diminish the noises in empirically observed correlation matrices.

A popular approach for the noise elimination of financial correlation matrices is the Random Matrix Theory, which studies the properties of matrix-form random variables—in particular, the density and behavior of eigenvalues. Its applications cover many of the fields of knowledge of recent years, such as statistical physics, dynamic systems, optimal control, and multivariate analysis.

Regarding applications in quantitative finance, Laloux et al. (1999) compared the empirical eigenvalues density of major stock market data with their theoretical prediction, assuming that the covariance matrix was random following a Wishart distribution (If a vector of random matrix variables follows a multivariate Gaussian distribution, then its Sample covariance matrix will follow a Wishart distribution (EDELMAN, 1988)).The results showed that over 94% of the eigenvalues fell within the theoretical bounds (defined in Edelman (1988)), implying that less than 6% of the eigenvalues contain actually useful information; moreover, the largest eigenvalue is significantly higher than the theoretical upper bound, which is evidence that the covariance matrix estimated via Markowitz is composed of few very informative principal components and many low-valued eigenvalues dominated by noise. Nobi et al. (2013) tested for the daily data of 20 global financial indexes from 2006 to 2011 and also found out that most eigenvalues fell into the theoretical range, suggesting a high presence of noises and few eigenvectors with very highly relevant information; particularly, this effect was even more prominent during a financial crisis. Although studies like Alaoui (2015) found a larger percentage of informative eigenvalues, the reported results show that the wide majority of principal components is still dominated by noisy information.

Plerou et al. (2002) found similar results, concluding that the top eigenvalues of the covariance matrices were stable in time and the distribution of their eigenvector components displayed systematic deviations from the Random Matrix Theory predicted thresholds. Furthermore, the paper pointed out that the top eigenvalues corresponded to an influence common to all stocks, representing the market's systematic risk, and their respective eigenvectors showed a prominent presence of central business sectors.

Sensoy, Yuksel and Erturk (2013) tested 87 benchmark financial indexes between

2009 and 2012, and also observed that the largest eigenvalue was more than 14 times larger than the Random Matrix Theory theoretical upper bound, while only less than 7% of the eigenvalues were larger than this threshold. Moreover, the paper identifies "central" elements that define the "global financial market" and analyzes the effects of the 2008 financial crisis in its volatility and correlation levels, concluding that the global market's dependence level generally increased after the crisis, thus making diversification less effective. Many other studies identified similar patterns in different financial markets and different time periods (REN; ZHOU, 2014; SHARMA; BANERJEE, 2015), evidencing the high levels of noise in correlation matrices and the relevance of filtering such noise for financial analysis. The effects of the covariance matrix cleaning using Random Matrix Theory in an emerging market was discussed in Eterovic and Eterovic (2013), which analyzed 83 stocks from the Chilean financial market between 2000 and 2011 and found out that the efficiency of portfolios generated using Markowitz (1952)'s model were largely improved.

Analogously, Eterovic (2016) analyzed the effects of covariance matrix filtering through the Random Matrix Theory using data from the stocks of the FTSE 100 Index between 2000 and 2012, confirming the distribution pattern of the eigenvalues of the covariance matrix, with the majority of principal components inside the bounds of the Marčenko-Pastur distribution, while the top eigenvalue was much larger than the remaining ones; in particular, the discrepancy of the top eigenvalue was even larger during the Crisis period. Moreover, Eterovic (2016) also found out that the performance improvement of the portfolios generated by a filtered covariance matrix filtering over a non-filtered one was strongly significant, evidencing the ability of the filtered covariance matrix to adapt to sudden volatility peaks.

Bouchaud and Potters (2009) summarized the potential applications of the Random Matrix Theory in financial problems, focusing on the cleaning of financial correlation matrices and the asymptotic behavior of its eigenvalues, whose density was enunciated in Marčenko and Pastur (1967)—and especially the largest one, which was described by the Tracy-Widom distribution (TRACY; WIDOM, 2002). The paper presents an empirical application using daily data of US stocks between 1993 and 2008, observing the correlation matrix of the 500 most liquid stocks in a sliding window of 1000 days with an interval of 100 days each, yielding 26 sample eigenvalue distributions. On average, the largest eigenvalue represents 21% of the sum of all eigenvalues. This is a stylized fact regarding the spectral properties of financial correlation matrices, as discussed in Akemann, Baik and Francesco (2011). Similar results were found in Conlon, Ruskin and Crane (2007), which analyzes the effects of "cleaning" the covariance matrix on better predictions of the risk of a portfolio, which may aid the investors to pick the best combination of hedge funds to avoid risk.

In financial applications, the covariance matrix is also important in multi-stage optimization problems, whose dimensionality often grows exponentially as the number of stages, financial assets or risk factor increase, thus demanding approximations using simulated scenarios to circumvent the curse of dimensionality (WAN; PEKNY; REKLAITIS, 2006). In this framework, an important requirement for the simulated scenarios is the absence of arbitrage opportunities, a condition which can be incorporated through resampling or increasing the number of scenarios (CONSIGLIO; CAROLLO; ZENIOS, 2016). Alternatively, (GEYER; HANKE; WEISSENSTEINER, 2014) defined three classes for arbitrage propensity and suggested a transformation on the covariance matrix's Cholesky decomposition that avoids the possibility of arbitrage in scenarios where it could theoretically exist. In this way, the application of the Random Matrix Theory on this method can improve the simulated scenarios in stochastic optimization problems, and consequently improve the quality of risk measurement and asset allocation decision-making.

Burda et al. (2004) provided a mathematical derivation of the relationship between the sample correlation matrix calculated using the conventional Pearson estimates with its population counterpart, discussing how the dependency structure of the spectral moments can be applied to filter out the noisy eigenvalues of the correlation matrix's spectrum. In fact, a reasonable choice of a 500×500 covariance matrix (like using the S&P 500 data for portfolio selection) induces a very high level of noise in addition to the signal that comes from the eigenvalues of the population covariance matrix; Laloux et al. (2000) used daily data of the S&P 500 between 1991 and 1996, and found out that the covariance matrix estimated by the classical Markowitz model highly underestimates the portfolio risks for a second time period (approximately three times lower than the actual values), a difference that is significantly lower for a cleaned correlation matrix, evidencing the high level of noise and the instability of the market dependency structure over time.

In view of the importance of controlling the complexity introduced alongside nonlinearities, in this paper we sought to verify whether the stylized behavior of the top eigenvalues persists after introducing nonlinearities into the covariance matrix, as well as the effect of cleaning the matrix's noises in the portfolio profitability and consistency over time, in order to obtain insights regarding the cost–benefit relationship between using higher degrees of nonlinearity to estimate the covariance between financial assets and the out-of-sample performance of the resulting portfolios.

4.3 Method

4.3.1 Mean-Variance Portfolio Optimization

Let $a_1, a_2, ..., a_p$ be the p available financial assets and r_{a_i} be the return vector of the i-th asset a_i, where the expected return vector and the covariance matrix are defined,

respectively, as $\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_p) = (\mathbb{E}[r_{a_1}], \mathbb{E}[r_{a_2}], ..., \mathbb{E}[r_{a_p}])$ and $\Sigma = (\sigma_{ij}), \quad i, j = 1, 2, ..., p$, with $\sigma_{ij} = cov(r_{a_i}, r_{a_j})$. Markowitz (1952)'s mean-variance portfolio optimization is basically a quadratic programming constrained optimization problem whose optimal solution $\boldsymbol{w} = (w_1, w_2, ..., w_p)^T, \quad \sum_{i=1}^{p} w_i = 1$ represents the weights allocated to each one of the p assets, such that the portfolio $\mathcal{P} = w_1 a_1 + w_2 a_2 + ... + w_p a_p$. Algebraically, the expected return and the variance of the resulting portfolio $\mathcal{P}$ are:

$$\begin{aligned} \mathbb{E}[\mathcal{P}] = & \quad \sum_{i=1}^{p} w_i \mathbb{E}[r_{a_i}] & = \boldsymbol{\mu}^T \boldsymbol{w} \in \mathbb{R} \\ \mathbb{V}[\mathcal{P}] = & \quad \sum_{i=1}^{p} \sum_{j=1}^{p} w_i w_j cov(r_{a_i}, r_{a_j}) & = \boldsymbol{w}^T \boldsymbol{\Sigma} \boldsymbol{w} \geq 0 \end{aligned}$$

With the non-allowance of a short selling constraint, the quadratic optimization problem is defined as:

$$\begin{aligned} \text{Minimize :} \quad & \frac{1}{2} \boldsymbol{w}^T \boldsymbol{\Sigma} \boldsymbol{w} \\ \text{Subject to :} \quad & \boldsymbol{\mu}^T \boldsymbol{w} = R, \boldsymbol{w}^T \mathbf{1} = 1, \boldsymbol{w} > \mathbf{0} \end{aligned} \tag{4.1}$$

which yields the weights that give away the less risky portfolio that provides an expected return equal to R; therefore, the portfolio $\mathcal{P}$ that lies on the efficient frontier for $\mathbb{E}[\mathcal{P}] = R$. The dual form of this problem has an analogous interpretation—instead of minimizing the risk at a given level of expected return, it maximizes the expected return given a certain level of tolerated risk.

Markowitz (1952)'s model is very intuitive, easy to interpret, and enjoys huge popularity to this very day, making it one of the main baseline models for portfolio selection. Moreover, it has only two inputs which are fairly easy to be estimated. Nevertheless, there are many different ways of doing so, which was the motivation of many studies to tackle this question, proposing alternative ways to estimate those inputs to find potentially better portfolios. The famous Black and Litterman (1992) model, for example, proposes a way to estimate the expected returns vector based on the combination of market equilibrium and the expectations of the investors operating in that market. In this paper, we focus on alternative ways to estimate the covariance matrix, and whether features like nonlinearities (Kernel functions) and noise filtering (Random Matrix Theory) can generate more profitable portfolio allocations.

4.3.2 Covariance Matrices

While Pearson's covariance estimator is consistent, studies like Huo, Kim and Kim (2012) pointed out that the estimates can be heavily influenced by outliers, which in turn leads to potentially suboptimal portfolio allocations. In this regard, the authors analyzed the effect of introducing robust estimation of covariance matrices, with the results of

the empirical experiments showing that the use of robust covariance matrices generated portfolios with larger profitabilities. Zhu, Welsch et al. (2018) found similar results, proposing a high-dimensional covariance estimator less prone to outliers and leading to more well-diversified portfolios, often with a higher alpha.

Bearing in mind the aforementioned findings of the literature, we tested KPCA and the noise filtering to many robust covariance estimators as well, in order to further investigate the effectiveness of nonlinearities introduction and the elimination of noisy eigenvalues to the portfolio's performance. Furthermore, we intended to check the relative effects of said improvements to Pearson and robust covariance matrices, and whether robust estimators remained superior under such conditions.

In addition to the Pearson covariance matrix $\mathbf{\Sigma} = \frac{1}{T}\sum_{i=1}^{p} \boldsymbol{x}_i \boldsymbol{x}_i^T$, where x_i is the return vector (centered in zero) of the i-th asset and T is the number of in-sample time periods, in this paper we considered four robust covariance estimators: the minimum covariance determinant (henceforth MCD) method (ROUSSEEUW, 1984), as estimated by the FASTMCD algorithm (ROUSSEEUW; DRIESSEN, 1999); the Reweighted MCD, following (HUBERT; DRIESSEN, 2004)'s algorithm; and the Orthogonalized Gnanadesikan-Kettenring (henceforth OGK) pairwise estimator (GNANADESIKAN; KETTENRING, 1972), following the algorithm of (MARONNA; ZAMAR, 2002).

The MCD method aims to find observations whose sample covariance has a minimum determinant, thus being less sensitive to non-persistent extreme events, such as an abrupt oscillation of price levels that briefly come back to normal. Cator and Lopuhaä (2012) demonstrated some statistical properties of this estimator, such as consistency and asymptotic convergence to the Gaussian distribution. The reweighted MCD estimator follows a similar idea, assigning weights to each observation and computing the covariance estimates based on the observations within a confidence interval, making the estimates even less sensitive to outliers and noisy datasets, as well as boosting the finite-sample efficiency of the estimator, as discussed in Croux and Haesbroeck (1999). Finally, the OGK approach takes univariate robust estimators of location and scale, constructing a covariance matrix based on those estimates and replacing the eigenvalues of that matrix with "robust variances", which are updated sequentially by weights based on a confidence interval cutoff.

4.3.3 Principal Component Analysis

Principal component analysis (henceforth PCA) is a technique for dimensionality reduction introduced by (PEARSON, 1901) which seeks to extract the important information from the data and to express this information as a set of new orthogonal variables called principal components, given that the independent variables of a dataset are generally correlated in some way. Each of these principal components is a linear combination

of the set of variables in which the coefficients show the importance of the variable to the component. By definition, the sum of all eigenvalues is equal to the total variance, as they represent an amount of observed information; therefore, each eigenvalue represents the variation explained of the i-th principal component PC_i, such that their values reflect the proportion of information maintained in the respective eigenvector, and thus are used to determine how many factors should be retained.

In a scenario with p independent variables, if it is assumed that the eigenvalues' distribution is uniform, then each eigenvalue would contribute to $\frac{1}{p}$ of the model's overall explanatory power. Therefore, taking a number $k < p$ of principal components that are able to explain more than $\frac{k}{p}$ of the total variance can be regarded as a "gain" in terms of useful information retaining and noise elimination. In the portfolio selection context, Kim and Jeong (2005) used PCA to decompose the correlation matrix of 135 stocks traded on the New York Stock Exchange (NYSE). Typically, the largest eigenvalue is considered to represent a market-wide effect that influences all stocks (DRIESSEN; MELENBERG; NIJMAN, 2003; PéRIGNON; SMITH; VILLA, 2007; BILLIO et al., 2012; ZHENG et al., 2012).

Consider $\boldsymbol{\Sigma}$ as a covariance matrix associated with the random vector $\boldsymbol{X} = [X_1, X_2..., X_p]$ with eigenvalues $\lambda_1 \geq \lambda_2... \geq \lambda_p \geq 0$, where the rotation of the axis in $\mathbb{R}^p$ yields the linear combinations:

$$
\begin{aligned}
Y_1 &= \boldsymbol{q}_1^T \boldsymbol{X} = q_{11}X_1 + q_{12}X_2 + ... + q_{1p}X_p \\
Y_2 &= \boldsymbol{q}_2^T \boldsymbol{X} = q_{21}X_2 + q_{22}X_2 + ... + q_{2p}X_p \\
&\vdots \\
Y_p &= \boldsymbol{q}_p^T \boldsymbol{X} = q_{p1}X_1 + q_{p2}X_2 + ... + q_{pp}X_p
\end{aligned}
$$

or

$$
\boldsymbol{Y} = \boldsymbol{Q}^T \boldsymbol{X}
$$

where Q_i are the eigenvectors from $\boldsymbol{\Sigma}$. Thus, the first principal component Y_1 is the projection in the direction in which the variance of the projection is maximized. So, we obtained $Y_1, Y_2...Y_p$ orthonormal vectors with maximum variability.

To obtain the associated eigenvectors, we solved for $\det(\boldsymbol{\Sigma} - \lambda \boldsymbol{I}) = 0$ to obtain the diagonal matrix composed by the eigenvalues. The variance of the i_{th} principal component of $\boldsymbol{\Sigma}$ is equal to its i-th eigenvalue λ_i. By construction, the principal component are pairwise orthogonal—that is, the covariance between the eigenvectors is $cov(Q_i\boldsymbol{X}, Q_j\boldsymbol{X}) = 0,\ i \neq j$. Algebraically, the i-th principal component Y_i can be obtained by solving the following expression for a_i (BENGTSSON; HOLST, 2002):

$$
\max_{q_i} \left\{ \frac{q_i \sum_{i=1}^{p} q_i}{\boldsymbol{q}_i^T \boldsymbol{q}_i} \qquad cov(Y_i, Y_j) = 0, \forall\ 0 < j < i \right\} \tag{4.2}
$$

In the field of dimensionality reduction, the interest in entropy, the entropy-based distance metric, has been investigated, where (JENSSEN, 2010) developed kernel entropy component analysis (KECA) for data transformation and dimensionality reduction, an extension of PCA mixture entropy and n dimensionality decomposition. (SHEKAR et al., 2011) shows that by using kernel entropy component analysis in an application on face recognition algorithm based on Renyi entropy component, certain eigenvalues and the corresponding eigenvectors will contribute more to the entropy estimate than others, since the terms depend on different eigenvalues and eigenvectors.

4.3.4 Kernel Principal Component Analysis and Random Matrix Theory

Let $\boldsymbol{X}$ be a $T \times p$ matrix, T being the observations, p the variables, and $\boldsymbol{\Sigma}$ the covariance matrix $p \times p$. The spectral decomposition of $\boldsymbol{\Sigma}$ is given by:

$$\lambda \boldsymbol{Q} = \boldsymbol{\Sigma} \boldsymbol{Q}$$

being $\lambda \geq 0$ the eigenvalues and $\boldsymbol{Q}$ the eigenvectors.

If the values of matrix X are random normalized values generated by a Gaussian distribution, then if $T \to \infty$ and $p \to \infty$ where $\Psi = \frac{T}{p} \geq 1$ the eigenvalues of matrix $\boldsymbol{\Sigma}$ result in the following probability density function (CONLON; RUSKIN; CRANE, 2007):

$$p(\lambda) = \frac{\Psi}{2\pi} \frac{\sqrt{(\lambda_{max} - \lambda)(\lambda - \lambda_{min})}}{\lambda} \tag{4.3}$$

where λ_{max} and λ_{min} are the bound given by:

$$\lambda_{min}^{max} = \left(1 + \frac{1}{\Psi} \pm 2\sqrt{\frac{1}{\Psi}}\right) \tag{4.4}$$

This result basically states that the eigenvalues of a purely random matrix based on distribution (4.3) tend to fall inside the theoretical boundaries; thus, eigenvalues larger than the upper bound are expected to contain useful information concerning an arbitrary matrix, whilst the noisy information is dispersed into the other eigenvalues, whose behavior is similar to the eigenvalues of a matrix with no information whatsoever.

There are many applications of the Random Matrix Theory (RMT) in the financial context. (BAI; SHI, 2011) used RMT to reduce the noise into data before to model the covariance matrix of assets on Asset Pricing Theory Models by using the Bayesian approach. The posteriori distribution was adjusted by Wishart Distribution using MCMC methods.

The procedures proposed by RMT for dispersion matrices noise filter in a finances context require careful use. The reasons are due to the "stylized facts" present in this

type of data as logarithmic transformations in the attempt for symmetric distributions of returns and the presence of extreme values. The work of (FRAHM; JAEKEL, 2005) deals with these problems and uses Tyler's robust M-estimator (TYLER, 1983) to estimate the dispersion matrix to then identify the non-random part with the relevant information via RMT using (MARČENKO; PASTUR, 1967) bounds.

The covariance matrix $\mathbf{\Sigma}$ can be factored as:

$$\mathbf{\Sigma} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{-1} \tag{4.5}$$

where $\mathbf{\Lambda}$ is a diagonal matrix composed by p eigenvalues $\lambda_i \geq 0$, $i = 1, 2, ..., p$ and each one of the p columns of $\boldsymbol{Q}$, $\boldsymbol{q}_i$, $i = 1, 2, ..., p$, are the eigenvectors associated with the i-th eigenvector λ_i. The idea is to perform the decomposition of $\mathbf{\Sigma}$ following Equation (4.5) and to filter out the eigenvalues which fall inside the boundaries postulated in Equation (4.4) and reconstruct $\mathbf{\Sigma}$ by multiplying back the filtered eigenvalue matrix to the eigenvector matrices, and then using the filtered matrix as input to Markowitz (1952)'s model.

Eigenvalues smaller than the upper bound of Equation (4.4) were considered as "noisy eigenvalues", while eigenvalues larger than the upper bound were considered "non-noisy". For the eigenvalue matrix filtering, we maintained all non-noisy eigenvalues and replaced all the remaining noisy ones by their average in order to preserve the stability (positive-definitiveness) and keep a fixed sum for the matrix's trace, following Sharifi et al. (2004) and Conlon, Ruskin and Crane (2007).

For eigenvalue matrix filtering, we maintained all non-noisy eigenvalues in Λ and replaced all the remaining noisy ones λ_i^{noise} by their average ($\bar{\lambda}_i^{noise}$):

$$\bar{\lambda}_i^{noise} = \sum_{i=1}^{\Omega \in noise} \frac{\lambda_i^{noise}}{\#\Omega \in noise}$$

After the filtering process, we multiplied back the filtered eigenvalue matrix to yield the "clean" covariance matrix:

$$\mathbf{\Sigma}^* = \boldsymbol{Q}\boldsymbol{\Lambda}^*\boldsymbol{Q}^{-1} \tag{4.6}$$

where $\mathbf{\Lambda}^*$ is a diagonal matrix composed of the cleaned eigenvalues.

The nonlinear estimation of the covariance matrix was achieved by means of a Kernel function, defined as:

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \varphi^T(\boldsymbol{x}_i) \cdot \varphi(\boldsymbol{x}_j) \in \mathbb{R}, \qquad i, j = 1, 2, ..., p \tag{4.7}$$

where $\varphi : \mathbb{R}^p \Rightarrow \mathbb{R}^q, p < q$ transforms the original data to a higher dimension, which can even be infinite, and the use of the kernel function prevents the need to explicitly

compute the functional form of $\varphi(\boldsymbol{x})$; instead, κ computes the inner product of φ. This is known as the *kernel trick*. The use of the Kernel function can circumvent the problem of high dimensionality induced by $\varphi(\boldsymbol{x})$ without the need to explicitly compute its functional form; instead, all nonlinear interactions between the original variables are synthesized in a real scalar. Since the inner product is a similarity measure in Hilbert spaces, the Kernel function can be seen as a way to measure the "margin" between the classes in high (or even infinite) dimensional spaces.

The following application of the Kernel function to the linearly estimated covariance matrix:

$$\boldsymbol{\Sigma} = \frac{1}{T}\sum_{i=1}^{p} \boldsymbol{x}_i \boldsymbol{x}_i^T \tag{4.8}$$

allows one to introduce a high number of nonlinear interactions in the original data and transform $\boldsymbol{\Sigma}$ into a Kernel covariance matrix:

$$\boldsymbol{\Sigma}_{\boldsymbol{\kappa}} = \frac{1}{T}\sum_{i=1}^{p} \varphi(\boldsymbol{x}_i)\varphi^T(\boldsymbol{x}_i) \tag{4.9}$$

In this paper, we tested the polynomial and Gaussian Kernels as κ. Both Kernels are widely used functions in the machine learning literature. The polynomial Kernel:

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = [(\boldsymbol{x}_i \cdot \boldsymbol{x}_j) + d]^q, d \in \mathbb{R}, q \in \mathbb{N}^+ \tag{4.10}$$

has a concise functional form, and is able to incorporate all cross-interactions between the explanatory variables that generate monomials with a degree less than or equal to a predefined q. This paper considered polynomial Kernels of degrees 2, 3, and 4 ($q = 2, 3, 4$). Note that the polynomial Kernel with $q = 1$ and $d = 0$ precisely yields the Pearson linear covariance matrix, so the polynomial Kernel covariance matrix is indeed a more general case of the former.

The Gaussian Kernel is the generalization of the polynomial Kernel for $q \to \infty$, and is one of the most widely used Kernels in machine learning literature. It enjoys huge popularity in various knowledge fields since this function is able to induce an infinite dimensional feature space while depending on only one scattering parameter σ. The expression of the Gaussian Kernel is given by:

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\boldsymbol{x}_i - \boldsymbol{x}_j^2}{2\sigma^2}\right), \sigma > 0 \tag{4.11}$$

The Kernel Principal Component Analysis (henceforth KPCA) is an extension of the linear PCA applied to the Kernel covariance matrix. Basically, the diagonalization problem returns linear combinations from the Kernel function's feature space $\mathbb{R}^q$, instead of the original input space $\mathbb{R}^p$ with the original variables. By performing the spectral decomposition in the Kernel covariance matrix:

$$\boldsymbol{\Sigma}_{\boldsymbol{\kappa}_{(pxp)}} = \boldsymbol{Q}\boldsymbol{\Lambda}_{\boldsymbol{\kappa}}\boldsymbol{Q}^{-1} \tag{4.12}$$

and extracting the largest eigenvalues of the Kernel covariance eigenvalue matrix $\mathbf{\Lambda_\kappa}$, we obtained the filtered Kernel covariance eigenvalue matrix $\mathbf{\Lambda^*_\kappa}$, which was then used to reconstruct the filtered Kernel covariance matrix:

$$\mathbf{\Sigma^*_{\kappa_{(pxp)}}} = \boldsymbol{Q\Lambda^*_\kappa Q^{-1}} \tag{4.13}$$

Finally, $\mathbf{\Sigma^*_\kappa}$ was used as an input for the Markowitz portfolio optimization model, and the resultant portfolio's profitability was compared to the portfolio generated by the linear covariance matrix and other aforementioned robust estimation methods, as well as their filtered counterparts. The analysis was reiterated for data from seven different markets, and the results are discussed in Section 4.5.

The pseudocode of our proposed approach is displayed as follows:

1. Estimate $\mathbf{\Sigma}$ for training set data;
2. Perform spectral decomposition of $\mathbf{\Sigma}$: $\mathbf{\Sigma = Q\Lambda Q^{-1}}$;
3. From the eigenvalues matrix $\mathbf{\Lambda}$, identify the noisy eigenvalues λ_i^{noise} based on the Random Matrix Theory upper bound;
4. Replace all noisy by their average: $\bar{\lambda}_i^{noise}$ to obtain the filtered eigenvalue matrix Λ^*;
5. Build the filtered covariance matrix $\boldsymbol{Q\Lambda^* Q^{-1}}$;
6. Use the filtered covariance matrix as input for Markowitz model and get the optimal portfolio weights from in-sample data;
7. Apply the in-sample optimal portfolio weights for out-of-sample data and obtain performance measures.

The above procedure was repeated for all seven datasets (NASDAQ 100, CAC 40, DAX-30, FTSE 100, NIKKEI 225, IBOVESPA, SSE 180). For Step **1** (estimation method of the covariance matrix), we applied eight different methods, namely: linear (Pearson), minimum covariance determinant (MCD), reweighted minimum covariance determinant (RMCD), Orthogonalized Gnanadesikan-Kettenring (OGK), Polynomial Kernel functions of degree 2 (K_POLY2), degree 3 (K_POLY3) and degree 4 (K_POLY4), and the Gaussian Kernel function (K_GAUSS).

4.4 Empirical Analysis

4.4.1 Performance Measures

The trade-off between risk and return has long been well-known in the finance literature, where higher expected return generally implies a greater level of risk, which

motivates the importance of considering risk-adjusted measures of performance. Therefore, it is not sufficient to view a portfolio's attractiveness only in terms of the cumulative returns that it offers, but instead, whether the return compensates for the level of risk that the allocation exposes the investor to. The Sharpe ratio (SHARPE, 1966) provides a simple way to do so.

Let $\mathcal{P}$ be a portfolio composed by a linear combination between assets whose expected return vector is $\boldsymbol{r}$, considering $\boldsymbol{w}$ as the weight vector of $\mathcal{P}$ and r_{f_t} as the risk-free rate at time t. Defining the mean excess return over the risk-free asset of $\mathcal{P}$ along the N out-of-sample time periods as:

$$\bar{\mu}_{\mathcal{P}} = \frac{1}{N}\sum_{t=1}^{N} \boldsymbol{w}_t^T \boldsymbol{r}_t - r_{f_t} \tag{4.14}$$

and defining the sample standard deviation of portfolio $\mathcal{P}$ as:

$$\sigma_{\mathcal{P}} = \sqrt{\frac{1}{N-1}\sum_{t=1}^{N} (\boldsymbol{w}_t^T \boldsymbol{r}_t - r_{f_t} - \bar{\mu}_{\mathcal{P}})^2} \tag{4.15}$$

The Sharpe ratio of portfolio $\mathcal{P}$ is given by:

$$ShR_{\mathcal{P}} = \frac{\bar{\mu}}{\sigma_{\mathcal{P}}} \tag{4.16}$$

While the Sharpe ratio gives a risk-adjusted performance measure for a portfolio and allows direct comparison between different allocations, it has the limitation of considering both the upside and the downside risks. That is, the uncertainty of profits is penalized in the Sharpe ratio expression, even though it is positive for an investor. Therefore, as discussed in works like Patton and Sheppard (2015) and Farago and Tédongap (2018), the decomposition of risk in "good variance" and "bad variance" can provide better asset allocation and volatility estimation, thus leading to better investment and risk management decisions. Therefore, instead of using the conventional standard deviation, which considers both methods of variance, Sortino and Price (1994) proposed an alternative performance measure that became known as the Sortino ratio, which balances the mean excess return only by the downside deviation. The Sortino ratio for portfolio $\mathcal{P}$ is given by:

$$SoR_{\mathcal{P}} = \frac{\bar{\mu}}{\sigma_{\mathcal{P}}^{-}} \tag{4.17}$$

where $\sigma_{\mathcal{P}}^{-}$ is the downside deviation:

$$\sigma_{\mathcal{P}}^{-} = \sqrt{\frac{1}{N-1}\sum_{t=1}^{N} min(0, (\boldsymbol{w}_t^T \boldsymbol{r}_t - r_{f_t} - \bar{\mu}_{\mathcal{P}})^2)} \tag{4.18}$$

Note that the downside deviation represents the standard deviation of negative portfolio returns, thus measuring only the "bad" side of volatility; for periods that the

portfolio yields a better return than the mean excess return over the risk-free asset, this upside deviation is not accounted for by the Sortino ratio.

Furthermore, we tested the statistical significance of the covariance matrix filtering improvement on the portfolio's performance. That is, instead of just comparing the values of the ratios, we tested to which extent the superiority of the noise-filtering approach was statistically significant. For each model and each analyzed market, we compared the Sharpe and Sortino ratios of the non-filtered covariance matrices with their respective filtered counterparts using Student's t tests. The null and alternative hypothesis are defined as follows:

$$\begin{cases} H_0: & ShR_{non-filtered} = ShR_{filtered} \\ H_A: & ShR_{non-filtered} < ShR_{filtered} \end{cases} \tag{4.19}$$

$$\begin{cases} H_0: & SoR_{non-filtered} = SoR_{filtered} \\ H_A: & SoR_{non-filtered} < SoR_{filtered} \end{cases} \tag{4.20}$$

Rejection of both null hypotheses implies that the Sharpe/Sortino ratio of the portfolio generated using the filtered covariance matrix is statistically larger than the portfolio yielded by the non-filtered matrix. The p-values for the hypothesis tests are displayed in tables 16 to 22.

4.4.2 Data

For the empirical application, we used data from seven markets—namely, the United States, United Kingdom, France, Germany, Japan, China, and Brazil; the chosen financial indexes representing each market were, respectively, NASDAQ-100, FTSE 100, CAC 40, DAX-30, NIKKEI 225, SSE 180 and Bovespa. We collected the daily return of the financial assets that composed those indexes during all time periods between 1 January 2000 and 16 August 2018, totaling 4858 observations for each asset. The data was collected from the Bloomberg terminal. The daily excess market return over the risk-free rate was collected from Kenneth R. French's data library.

We split the datasets into two mutually exclusive subsets: we allocated the observations between 1 January 2000 and 3 November 2015 (85% of the whole dataset, 4131 observations) for the training (in-sample) subset and the observations between 4 November 2015 and 16 August 2018 (the remaining 15%, 727 observations) for the test (out-of-sample) subset. For each financial market and each covariance matrix method, we estimated the optimal portfolio for the training subset and applied the optimal weights for the test subset data. The cumulative return of the portfolio in the out-of-sample periods, their Sharpe and Sortino ratios, information regarding the non-noisy eigenvalues and p-values of tests (4.19) and (4.20) are displayed in Tables 16 to 22.

4.5 Results and Discussion

The cumulative returns and risk-adjusted performance metrics are presented in Tables 16–22, as well as information regarding the non-noisy eigenvalues and the p-values of the hypothesis tests. Figures 6–12 show the improvement of filtered covariance matrices over their non-filtered counterparts for each market and estimation method. The results are summarized as follows:

Covariance matrix	Method	CR (%)	Sharpe ratio	Sortino ratio	λ^*	$\lambda^*_{variance}(\%)$	λ^{top}	$\lambda^{top}_{variance}(\%)$	p_{Sharpe}	$p_{Sortino}$
Non-filtered	Pearson	22.3297	0.3252	0.4439						
	MCD	19.1094	0.2713	0.3690						
	RMCD	18.6733	0.2632	0.3574						
	OGK	21.2332	0.3037	0.4138						
	K_POLY2	28.7582	0.3808	0.5144						
	K_POLY3	28.7561	0.3884	0.5253						
	K_POLY4	29.7912	0.4108	0.5561						
	K_GAUSS	13.7226	0.1703	0.2304						
Filtered	Pearson	18.9984	0.2834	0.3847	5	45.38%	20.0680	33.33%	0.9432	0.9874
	MCD	23.9648	0.3595	0.4924	5	51.1%	24.6837	40.99%	0.0004	$< 10^{-4}$
	RMCD	23.4073	0.3459	0.4730	5	51.19%	24.6470	40.93%	0.0011	$< 10^{-4}$
	OGK	23.6193	0.3512	0.4809	5	49.53%	23.7152	39.39%	0.0382	0.0061
	K_POLY2	15.831	0.2218	0.3015	5	38.24%	16.1131	26.76%	> 0.9999	> 0.9999
	K_POLY3	16.7263	0.2496	0.3389	5	26.23%	9.2748	15.4%	> 0.9999	> 0.9999
	K_POLY4	16.186	0.2417	0.3283	5	19.29%	5.7377	9.53%	> 0.9999	> 0.9999
	K_GAUSS	21.823	0.2496	0.3435	5	67.89%	24.9393	41.42%	0.0015	$< 10^{-4}$

Table 16 – Summary results for assets of NASDAQ-100 Index: CR is the cumulative return of the optimal portfolio in the out-of-sample period; λ^* is the number of non-noisy eigenvalues of the respective covariance matrix; $\lambda^*_{variance}(\%)$ is the percentage of variance explained by the non-noisy eigenvalues; λ^{top} is the value of the top eigenvalue; $\lambda^{top}_{variance}(\%)$ is the percentage of variance that the top eigenvalue is responsible for; p_{Sharpe} is the p-value of the hypothesis test **4.19**; and $p_{Sortino}$ is the p-value of the hypothesis test **4.20**.

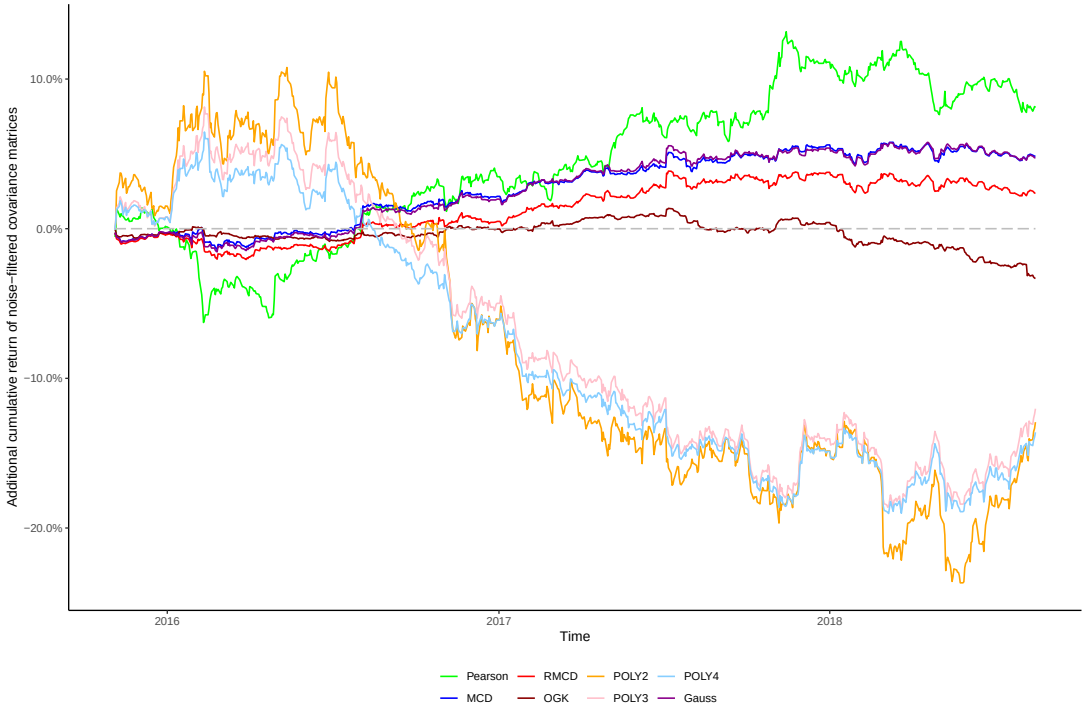

Figure 6 – Cumulative return improvement of noise-filtered covariance matrices over non-filtered ones for assets of NASDAQ-100 Index during the out-of-sample period.

Covariance matrix	Method	CR (%)	Sharpe ratio	Sortino ratio	λ^*	$\lambda^*_{variance}(\%)$	λ^{top}	$\lambda^{top}_{variance}(\%)$	p_{Sharpe}	$p_{Sortino}$
Non-filtered	Pearson	-16.8525	-0.2443	-0.3236						
	MCD	-23.9938	-0.3252	-0.4203						
	RMCD	-24.2595	-0.3272	-0.4223						
	OGK	-23.5119	-0.3223	-0.4178						
	K_POLY2	-2.4443	-0.0377	-0.0483						
	K_POLY3	-3.0975	-0.0453	-0.0575						
	K_POLY4	-3.1496	-0.0462	-0.0583						
	K_GAUSS	-5.4357	-0.0772	-0.1022						
Filtered	Pearson	-15.1099	-0.2246	-0.2986	6	52.52%	22.7137	38.24%	0.0222	0.0051
	MCD	-22.5761	-0.3148	-0.4096	6	55.87%	25.6111	43.12%	0.1547	0.1491
	RMCD	-22.8926	-0.3178	-0.4131	6	56.27%	25.8719	43.55%	0.1813	0.1852
	OGK	-22.3237	-0.3142	-0.4104	6	55.15%	25.2449	42.5%	0.2137	0.2326
	K_POLY2	-13.825	-0.2029	-0.2711	5	47.84%	21.2488	35.77%	> 0.9999	> 0.9999
	K_POLY3	-12.2619	-0.1812	-0.2413	7	38.27%	13.3597	22.49%	> 0.9999	> 0.9999
	K_POLY4	-10.2092	-0.1539	-0.2028	9	33.23%	8.6809	14.61%	> 0.9999	> 0.9999
	K_GAUSS	6.9977	0.0657	0.0908	7	75.37%	25.9374	43.66%	$< 10^{-4}$	$< 10^{-4}$

Table 17 – Summary results for assets of FTSE 100 Index: CR is the cumulative return of the optimal portfolio in the out-of-sample period; λ^* is the number of non-noisy eigenvalues of the respective covariance matrix; $\lambda^*_{variance}(\%)$ is the percentage of variance explained by the non-noisy eigenvalues; λ^{top} is the value of the top eigenvalue; $\lambda^{top}_{variance}(\%)$ is the percentage of variance that the top eigenvalue is responsible for; p_{Sharpe} is the p-value of the hypothesis test **4.19**; and $p_{Sortino}$ is the p-value of the hypothesis test **4.20**.

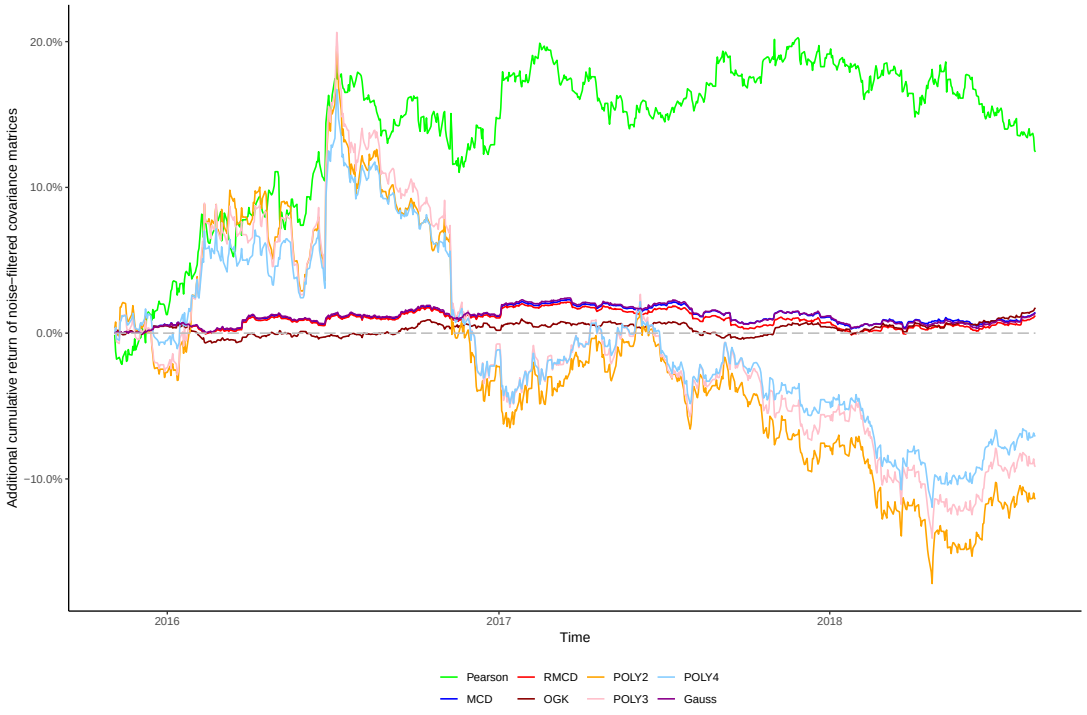

Figure 7 – Cumulative return improvement of noise-filtered covariance matrices over non-filtered ones for assets of FTSE 100 Index during the out-of-sample period.

Covariance matrix	Method	CR (%)	Sharpe ratio	Sortino ratio	λ^*	$\lambda^*_{variance}(\%)$	λ^{top}	$\lambda^{top}_{variance}(\%)$	p_{Sharpe}	$p_{Sortino}$
Non-filtered	Pearson	16.2333	0.2015	0.2882						
	MCD	17.2074	0.2182	0.3117						
	RMCD	17.4111	0.2216	0.3165						
	OGK	17.6784	0.2264	0.3235						
	K_POLY2	11.8756	0.1423	0.1963						
	K_POLY3	10.6055	0.1311	0.1793						
	K_POLY4	9.5146	0.1188	0.1614						
	K_GAUSS	12.3998	0.1348	0.1928						
Filtered	Pearson	17.4651	0.2238	0.3199	3	56.82%	14.1697	48.52%	0.0147	0.0010
	MCD	18.9068	0.2475	0.3533	2	58.57%	15.9837	54.73%	0.0022	$< 10^{-4}$
	RMCD	19.0796	0.2504	0.3575	2	58.38%	15.9013	54.45%	0.0019	$< 10^{-4}$
	OGK	18.6063	0.2423	0.3461	2	56.89%	15.4144	52.78%	0.0578	0.0126
	K_POLY2	16.5982	0.2076	0.2969	3	51.5%	12.5296	42.9%	$< 10^{-4}$	$< 10^{-4}$
	K_POLY3	17.8811	0.2289	0.3274	4	42.31%	8.6342	29.57%	$< 10^{-4}$	$< 10^{-4}$
	K_POLY4	17.7003	0.2333	0.3311	4	33.88%	6.1270	20.98%	$< 10^{-4}$	$< 10^{-4}$
	K_GAUSS	11.5206	0.1228	0.1757	4	78.74%	16.0889	55.09%	0.8828	0.9549

Table 18 – Summary results for assets of CAC 40 Index: CR is the cumulative return of the optimal portfolio in the out-of-sample period; λ^* is the number of non-noisy eigenvalues of the respective covariance matrix; $\lambda^*_{variance}(\%)$ is the percentage of variance explained by the non-noisy eigenvalues; λ^{top} is the value of the top eigenvalue; $\lambda^{top}_{variance}(\%)$ is the percentage of variance that the top eigenvalue is responsible for; p_{Sharpe} is the p-value of the hypothesis test **4.19**; and $p_{Sortino}$ is the p-value of the hypothesis test **4.20**.

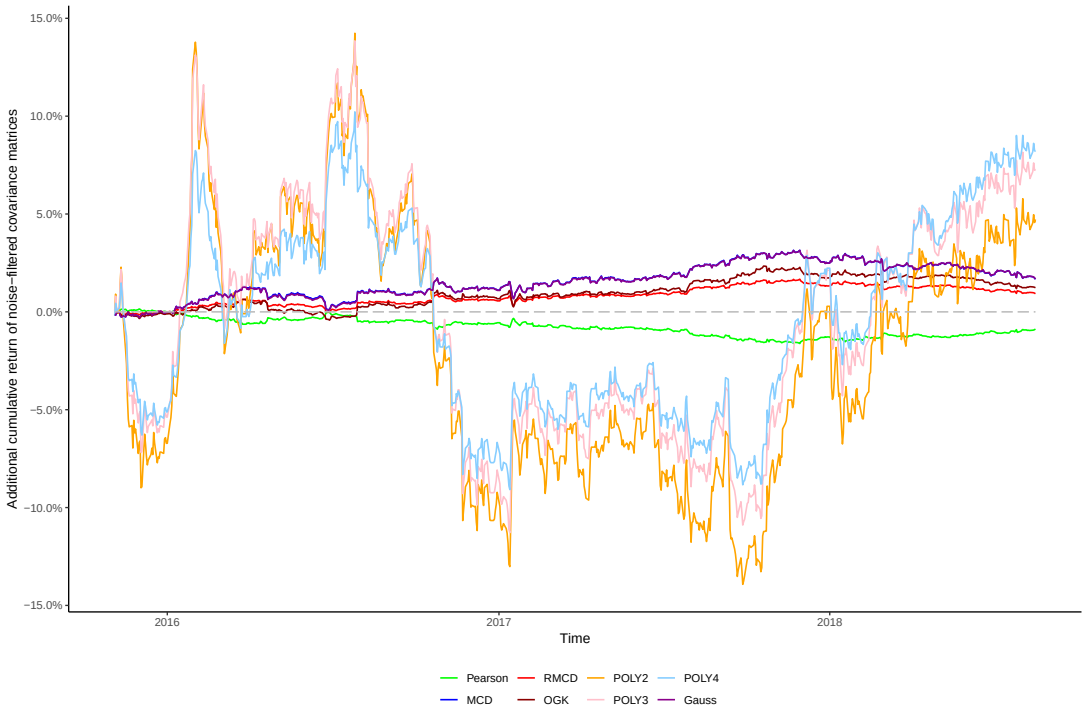

Figure 8 – Cumulative return improvement of noise-filtered covariance matrices over non-filtered ones for assets of CAC 40 Index during the out-of-sample period.

Covariance matrix	Method	CR (%)	Sharpe ratio	Sortino ratio	λ^*	$\lambda^*_{variance}(\%)$	λ^{top}	$\lambda^{top}_{variance}(\%)$	p_{Sharpe}	$p_{Sortino}$
Non-filtered	Pearson	6.3447	0.0772	0.1027						
	MCD	-1.5643	-0.0315	-0.0414						
	RMCD	-0.378	-0.0161	-0.0212						
	OGK	5.3011	0.0615	0.0815						
	K_POLY2	-4.6104	-0.0733	-0.0949						
	K_POLY3	-0.6555	-0.0204	-0.0265						
	K_POLY4	1.7874	0.0131	0.0171						
	K_GAUSS	-10.2399	-0.1311	-0.1720						
Filtered	Pearson	10.2332	0.1346	0.1796	3	55.24%	11.0402	46.1%	0.0014	$< 10^{-4}$
	MCD	7.0445	0.0866	0.1149	2	58.39%	12.8292	53.57%	$< 10^{-4}$	$< 10^{-4}$
	RMCD	7.5928	0.0942	0.1254	2	58.88%	12.9601	54.11%	$< 10^{-4}$	$< 10^{-4}$
	OGK	9.8916	0.1286	0.1715	2	56.32%	12.3346	51.5%	0.0003	$< 10^{-4}$
	K_POLY2	4.3642	0.0484	0.0640	2	46.78%	9.9835	41.69%	$< 10^{-4}$	$< 10^{-4}$
	K_POLY3	6.7303	0.0830	0.1099	3	38.77%	6.9275	28.93%	$< 10^{-4}$	$< 10^{-4}$
	K_POLY4	9.7678	0.1297	0.1717	4	35.17%	5.0114	20.93%	$< 10^{-4}$	$< 10^{-4}$
	K_GAUSS	-18.5834	-0.2365	-0.3050	2	71.04%	13.7234	57.3%	> 0.9999	> 0.9999

Table 19 – Summary results for assets of DAX-30 Index: CR is the cumulative return of the optimal portfolio in the out-of-sample period; λ^* is the number of non-noisy eigenvalues of the respective covariance matrix; $\lambda^*_{variance}(\%)$ is the percentage of variance explained by the non-noisy eigenvalues; λ^{top} is the value of the top eigenvalue; $\lambda^{top}_{variance}(\%)$ is the percentage of variance that the top eigenvalue is responsible for; p_{Sharpe} is the p-value of the hypothesis test **4.19**; and $p_{Sortino}$ is the p-value of the hypothesis test **4.20**.

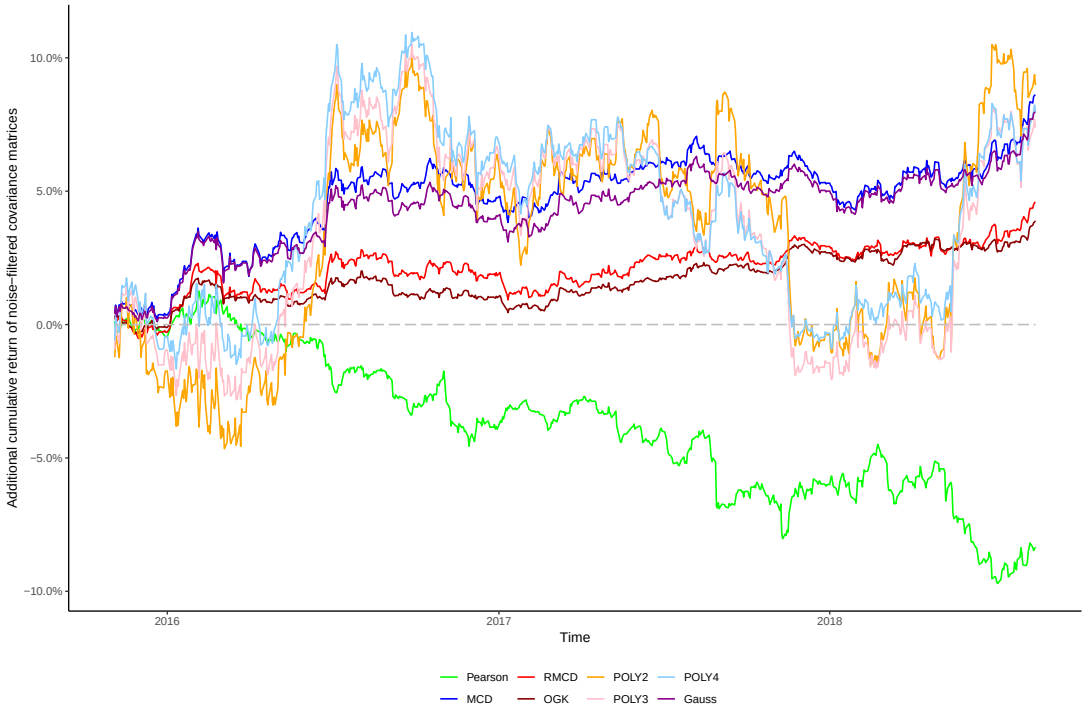

Figure 9 – Cumulative return improvement of noise-filtered covariance matrices over non-filtered ones for assets of the DAX-30 Index during the out-of-sample period.

Covariance matrix	Method	CR (%)	Sharpe ratio	Sortino ratio	λ^*	$\lambda^*_{variance}(\%)$	λ^{top}	$\lambda^{top}_{variance}(\%)$	p_{Sharpe}	$p_{Sortino}$
Non-filtered	Pearson	19.0365	0.2104	0.2976						
	MCD	17.9163	0.1979	0.2791						
	RMCD	18.3996	0.1983	0.2803						
	OGK	17.833	0.1951	0.2757						
	K_POLY2	8.5753	0.0959	0.1325						
	K_POLY3	10.6699	0.1233	0.1700						
	K_POLY4	13.1313	0.1553	0.2145						
	K_GAUSS	14.5078	0.1586	0.2236						
Filtered	Pearson	19.4964	0.2231	0.3161	12	54.88%	57.4396	39.38%	0.1347	0.0540
	MCD	18.266	0.2025	0.2855	11	57.24%	63.4158	43.48%	0.3498	0.2938
	RMCD	19.0273	0.2119	0.2987	12	58.83%	65.3846	44.83%	0.1235	0.0591
	OGK	19.0061	0.2142	0.3023	11	56.5%	62.0915	42.57%	0.0501	0.0111
	K_POLY2	15.1032	0.1637	0.2314	11	47.71%	49.6729	34.06%	$< 10^{-4}$	$< 10^{-4}$
	K_POLY3	16.8414	0.1890	0.2661	13	35.62%	30.0585	20.61%	$< 10^{-4}$	$< 10^{-4}$
	K_POLY4	18.2374	0.2090	0.2943	14	27.44%	18.6121	12.76%	$< 10^{-4}$	$< 10^{-4}$
	K_GAUSS	12.6904	0.1385	0.1953	15	72.24%	42.7789	29.33%	0.9570	0.9923

Table 20 – Summary results for assets of NIKKEI 225 Index: CR is the cumulative return of the optimal portfolio in the out-of-sample period; λ^* is the number of non-noisy eigenvalues of the respective covariance matrix; $\lambda^*_{variance}(\%)$ is the percentage of variance explained by the non-noisy eigenvalues; λ^{top} is the value of the top eigenvalue; $\lambda^{top}_{variance}(\%)$ is the percentage of variance that the top eigenvalue is responsible for; p_{Sharpe} is the p-value of the hypothesis test **4.19**; and $p_{Sortino}$ is the p-value of the hypothesis test **4.20**.

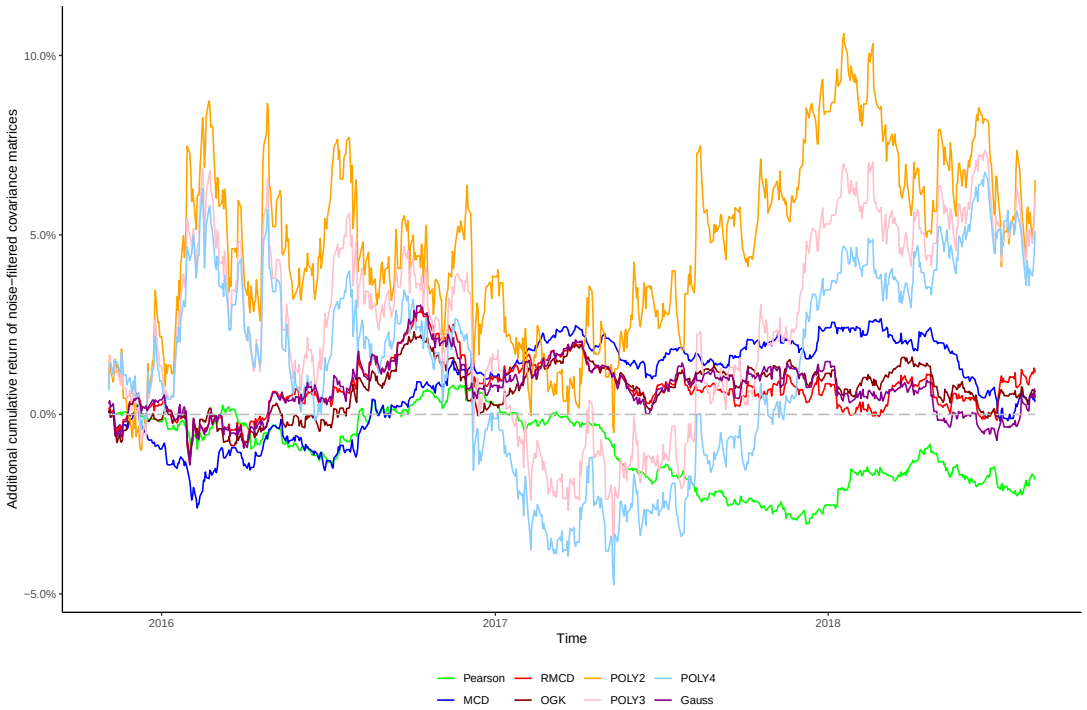

Figure 10 – Cumulative return improvement of noise-filtered covariance matrices over non-filtered ones for assets of the NIKKEI 225 Index during the out-of-sample period.

Covariance matrix	Method	CR (%)	Sharpe ratio	Sortino ratio	λ^*	$\lambda^*_{variance}(\%)$	λ^{top}	$\lambda^{top}_{variance}(\%)$	p_{Sharpe}	$p_{Sortino}$
Non-filtered	Pearson	-24.4861	-0.2945	-0.3765						
	MCD	-18.4543	-0.2139	-0.2762						
	RMCD	-20.8369	-0.2393	-0.3073						
	OGK	-22.9376	-0.2617	-0.3364						
	K_POLY2	-36.7953	-0.3531	-0.4459						
	K_POLY3	-35.2879	-0.3460	-0.4335						
	K_POLY4	-34.3716	-0.3422	-0.4258						
	K_GAUSS	-33.6337	-0.3735	-0.4744						
Filtered	Pearson	-21.0991	-0.2587	-0.3308	11	50.96%	56.5957	38.99%	0.0011	$< 10^{-4}$
	MCD	-25.1805	-0.2913	-0.3724	11	49.85%	54.7101	37.69%	> 0.9999	> 0.9999
	RMCD	-20.685	-0.2379	-0.3053	11	50.78%	56.5502	38.96%	0.4543	0.4344
	OGK	-21.7307	-0.2520	-0.3235	11	48.66%	52.5361	36.2%	0.2154	0.1482
	K_POLY2	-26.5935	-0.3140	-0.3978	12	41.25%	42.7236	29.44%	0.0007	$< 10^{-4}$
	K_POLY3	-28.6612	-0.3292	-0.4140	13	28.83%	24.2135	16.68%	0.0870	0.0565
	K_POLY4	-28.9269	-0.3338	-0.4186	12	20.18%	14.1161	9.73%	0.2469	0.2801
	K_GAUSS	-38.4531	-0.4102	-0.5175	12	69.52%	60.1106	41.42%	0.9986	0.9998

Table 21 – Summary results for assets of SSE 180 Index: CR is the cumulative return of the optimal portfolio in the out-of-sample period; λ^* is the number of non-noisy eigenvalues of the respective covariance matrix; $\lambda^*_{variance}(\%)$ is the percentage of variance explained by the non-noisy eigenvalues; λ^{top} is the value of the top eigenvalue; $\lambda^{top}_{variance}(\%)$ is the percentage of variance that the top eigenvalue is responsible for; p_{Sharpe} is the p-value of the hypothesis test **4.19**; and $p_{Sortino}$ is the p-value of the hypothesis test **4.20**.

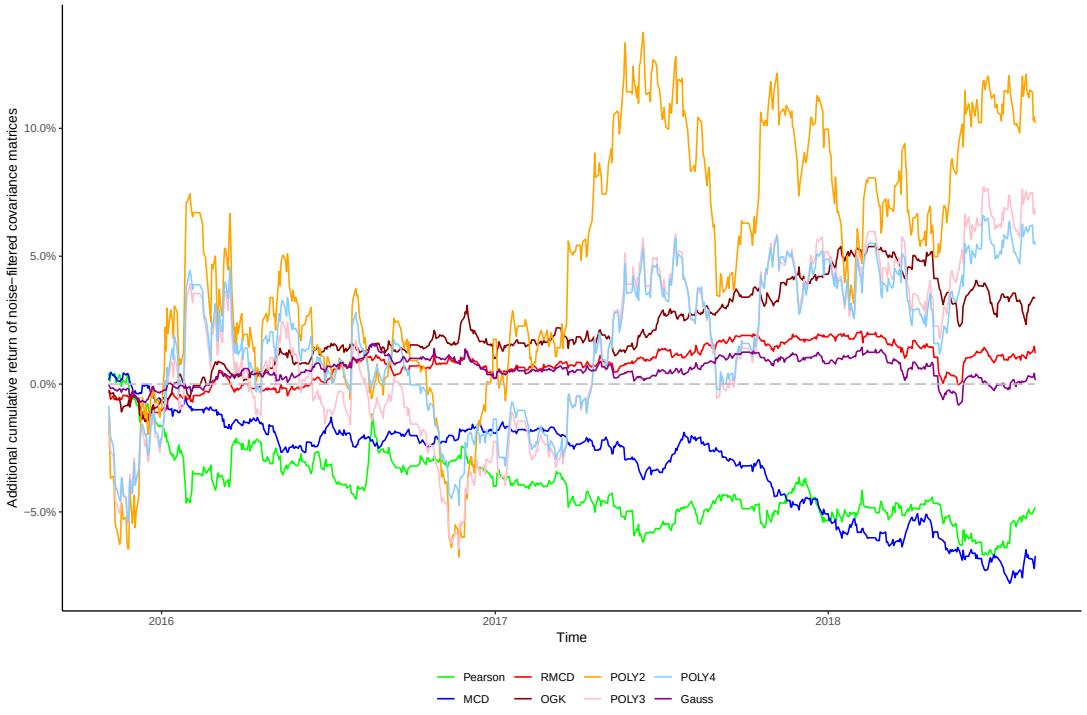

Figure 11 – Cumulative return improvement of noise-filtered covariance matrices over non-filtered ones for assets of SSE 180 Index during the out-of-sample period.

Covariance matrix	Method	CR (%)	Sharpe ratio	Sortino ratio	λ^*	$\lambda^*_{variance}(\%)$	λ^{top}	$\lambda^{top}_{variance}(\%)$	p_{Sharpe}	$p_{Sortino}$
Non-filtered	Pearson	9.3348	0.0636	0.0871						
	MCD	3.4975	0.0206	0.0280						
	RMCD	1.8602	0.0079	0.0107						
	OGK	3.0337	0.0167	0.0227						
	K_POLY2	15.2198	0.1127	0.1521						
	K_POLY3	16.2334	0.1184	0.1594						
	K_POLY4	16.6977	0.1194	0.1605						
	K_GAUSS	32.0362	0.1934	0.2657						
Filtered	Pearson	-3.5439	-0.0334	-0.0453	2	58.59%	13.5231	54.46%	> 0.9999	> 0.9999
	MCD	-3.8358	-0.0364	-0.0492	2	55.01%	12.5411	50.51%	0.9994	> 0.9999
	RMCD	-1.6626	-0.0191	-0.0258	2	54.11%	12.2963	49.52%	0.9329	0.9787
	OGK	-4.5348	-0.0412	-0.0557	2	54.81%	12.5097	50.38%	0.9994	> 0.9999
	K_POLY2	3.7777	0.0217	0.0296	2	47.88%	10.6994	43.09%	> 0.9999	> 0.9999
	K_POLY3	-4.0389	-0.0370	-0.0499	4	43.39%	7.3663	29.67%	> 0.9999	> 0.9999
	K_POLY4	-9.6085	-0.0809	-0.1087	4	35.63%	5.2703	21.23%	> 0.9999	> 0.9999
	K_GAUSS	31.7689	0.1916	0.2631	2	77.51%	16.0176	64.51%	0.5383	0.5568

Table 22 – Summary results for assets of Bovespa Index: CR is the cumulative return of the optimal portfolio in the out-of-sample period; λ^* is the number of non-noisy eigenvalues of the respective covariance matrix; $\lambda^*_{variance}(\%)$ is the percentage of variance explained by the non-noisy eigenvalues; λ^{top} is the value of the top eigenvalue; $\lambda^{top}_{variance}(\%)$ is the percentage of variance that the top eigenvalue is responsible for; p_{Sharpe} is the p-value of the hypothesis test **4.19**; and $p_{Sortino}$ is the p-value of the hypothesis test **4.20**.

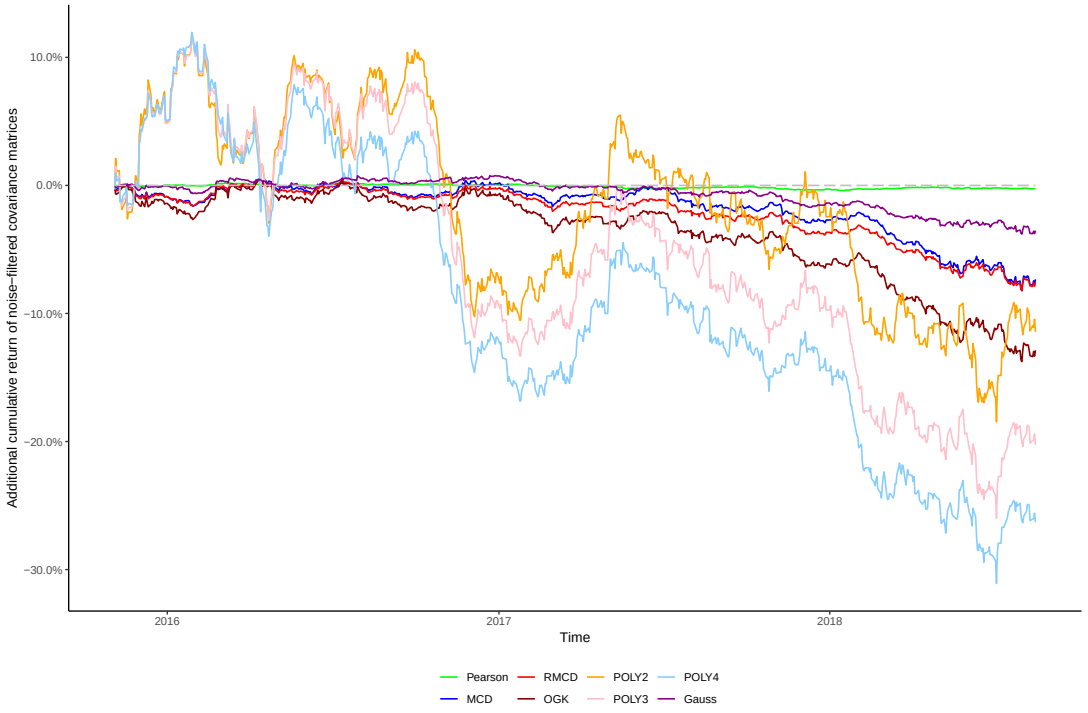

Figure 12 – Cumulative return improvement of noise-filtered covariance matrices over non-filtered ones for assets of Bovespa 100 Index during the out-of-sample period.

For the non-filtered covariance matrices, the overall performance of the linear Pearson estimates was better than robust estimation methods in most markets, although it was outperformed by all three robust methods (MCD, RMCD, and OGK) for the

CAC and SSE indexes. In comparison to the nonlinear covariance matrices induced by the application of Kernel functions, the linear approaches performed better in four out of the seven analyzed markets (CAC, DAX, NIKKEI, and SSE), although in the other three markets the nonlinear models performed better by a fairly large margin. Between the robust estimators, the performance results were similar, slightly favoring the OGK approach. Amongst the nonlinear models, the Gaussian Kernel generally performed worse than the polynomial Kernels—an expected result, as the Gaussian Kernel incorporates polynomial interactions that effectively tends to infinity-degree, which naturally inserts a large amount of noisy information; the only market where the Gaussian Kernel performed notably better was the Brazilian one, which is considered to be an "emerging economy" and a less efficient market compared to the United States or Europe; even though Brazil is the leading market in Latin America, this market's liquidity, transaction flows, and informational efficiency are quite smaller compared to major financial markets (For a broad discussion about the dynamics of financial markets of emerging economies, see Karolyi (2015)). Therefore, it is to be expected that such a market contains more levels of "noise", such that a function that incorporates a wider range of nonlinear interactions tend to perform better.

As for the filtered covariance matrices, the Pearson estimator and the robust estimators showed similar results, with no major overall differences in profitability or risk-adjusted measures—Pearson performed worst than MCD, RMCD, and OGK for NASDAQ and better for FTSE and DAX. In comparison to MCD and OGK, the RMCD showed slightly better performance. Similarly to the non-filtered cases, the polynomial Kernels yielded generally better portfolios in most markets. Concerning the Gaussian Kernel, even though its filtered covariance matrix performed particularly well for FTSE and Bovespa, it showed very bad results for the German and Chinese markets, suggesting that an excessive introduction of nonlinearities may bring along more costs than improvements. Nevertheless, during the out-of-sample periods, the British and Brazilian markets underwent exogenous events—namely the effects of the "Brexit" referendum for the United Kingdom and the advancements of the "Car Wash" (*Lava Jato*) operation, which led to events like the prison of Eduardo Cunha (former President of the Chamber of Deputies) in October 2016; and Luis Inácio da Silva (former President of Brazil) in April 2018—that may have affected their respective systematic levels of risk and profitability, potentially compromising the market as a whole. In this sense, the fact that the Gaussian Kernel-filtered covariance matrices in those markets performed better than the polynomial Kernels is evidence that the additional levels of "complexity" in those markets may demand the introduction of more complex nonlinear interactions to make good portfolio allocations. These results are also consistent with the finding of Sandoval-Jr, Bortoluzzo and Venezuela (2014), which pointed out that covariance matrix cleaning may actually lead to the worst portfolio performances in periods of high volatility.

Regarding the principal components of the covariance matrices and the dominance of the top eigenvalue discussed by the literature, the results showed that for all models and markets, the first eigenvalue of the covariance matrix was much bigger than the theoretical bound λ_{max}, which is consistent with the stylized facts discussed in Section 4.2. Moreover, for the vast majority of the cases (44 out of 54), the single top eigenvalue λ^{top} contained more than 25% of all the variance. This result is consistent with the finding of previous similar works stated in the literature review section (Sensoy, Yuksel and Erturk (2013) and others): the fact that a single principal component concentrated over 25% of the information is evidence that it captures the systematic risk, the very slice of the risk which cannot be diversified—in other words, the share of the risk that persists, regardless of the weight allocation. The results persisted for the eigenvalues above the upper bound of Equation (4.4): in more than half of the cases (31 out of 54), the "non-noisy" eigenvalues represented more than half of the total variance. The concentration of information in non-noisy eigenvalues in polynomial Kernels was weaker than the linear covariance matrices, while for the Gaussian Kernel the percentage of variance retained was even larger—around 70% of the total variance for all seven markets.

Finally, the columns p_{Sharpe} and $p_{Sortino}$ show the statistical significance of the improvement of Sharpe and Sortino ratios brought about by the introduction of noise filtering based on the Random Matrix Theory. The results indicate that, while in some cases the noise filtering worked very well, in other cases it actually worsened the portfolio's performances. Therefore, there is evidence that better portfolios can be achieved by eliminating the "noisy eigenvalues", but the upper bound given by Equation (4.4) may be classifying actually informative principal components as "noise". Especially concerning Kernel covariance matrices, the effects of the eigenvalues cleaning seemed unstable, working well in some cases and very bad in others, suggesting that the dynamics of the eigenvalues in nonlinear covariance matrices follow a different dynamic than linear ones, and the information that is considered to be "noise" for linear estimates can actually be informative in nonlinear domains. At the usual 95% confidence level, evidences of statistical superiority of filtered covariance matrices was present in 25 out of 54 cases for the Sharpe ratio (rejection of null hypothesis in (4.19)) and 26 out of 54 for the Sortino ratio (rejection of null hypothesis in (4.20)). The markets in which more models showed significant improvement using the Random Matrix Theory were the French and the German; on the other hand, again, for a less efficient financial market like the Brazilian one, the elimination of noisy eigenvalues yielded the worst performances (the profitability of all portfolios actually dropped), again consistent with the finding of Sandoval-Jr, Bortoluzzo and Venezuela (2014).

4.6 Conclusions

In this paper, the effectiveness of introducing nonlinear interactions to the covariance matrix estimation and its noise filtering using the Random Matrix Theory was tested with daily data from seven different financial markets. We tested eight estimators for the covariance matrix and evaluated the statistical significance of the noise-filtering improvement on portfolio performance. While the cleaning of noisy eigenvalues did not show significant improvements in every analyzed market, the out-of-sample Sharpe and Sortino ratios of the portfolios were significantly improved for almost half of all tested cases. The findings of this paper can potentially aid the investment decision for scholars and financial market participants, as well as providing both theoretical and empirical tools for the construction of more profitable and less risky trading strategies, as well as exploring potential weaknesses of traditional linear methods of covariance estimation.

We also tested the introduction of different degrees of nonlinearities to the covariance matrices by means of Kernel functions, with varied results: while in some cases, the Kernel approach managed to get better results, for others the addition yielded a much worse performance, indicating that the use of Kernels represent a high boost of the models' complexity levels, which are not always compensated by better asset allocations, even when part of the said additional complexity is filtered out. This implies that the noise introduced by nonlinear features can surpass the additional predictive power which they aggregate to the Markowitz model. To further investigate this result, future developments include testing other Kernel functions besides the polynomial and the Gaussian to investigate whether alternative frameworks of nonlinear dependence can show better results. For instance, the results shown by different classes of Kernel functions (GENTON, 2001) may fit better into the financial markets' stylized facts and reveal underlying patterns based on the Kernel's definition. Tuning the hyperparameters for each Kernel can also influence the model's performance decisively.

Although the past performance of a financial asset does not determine its future performance, in this paper we kept in the dataset only the assets that composed of the seven financial indexes during the whole period between 2000 and 2018, thus not considering the possible survivorship bias in the choice of the assets which can affect the model's implications (BROWN; GOETZMANN; ROSS, 1992). As for future advancements, the difference between the "surviving" assets from the others can be analyzed as well. Other potential improvements include the replication of the analysis for other financial indexes or markets and other time periods, incorporation of transaction costs, and comparison with other portfolio selection models apart from Markowitz's.

This paper focused on the introduction on nonlinear interactions to the covariance matrix estimation. Thus, a limitation was the choice of the filtering methods, as the replacement procedure that we adopted was not the only one that the literature on

the Random Matrix Theory recommends. Alternative filtering methods documented by studies like Guhr and Kälber (2003) and Daly, Crane and Ruskin (2008), such as exponential weighting and Krzanowski stability maximization, may allow for better modeling of underlying patterns of financial covariance structures and also lead to better portfolio allocations, such that the application of those methods and comparison to our proposed methods can be a subject of future research in this agenda.

5 Does all learning lead to the efficiency? Deep neural networks, feature selection and technical analysis indicators in stock price direction forecasting

Abstract

This paper analyzes the performance of deep neural network models to predict the stock price movement of financial assets based on technical analysis indicators used as explanatory variables in the recent Literature and specialized trading websites. We applied three feature selection methods to shrink the feature set and seeking to eliminate redundant information from similar indicators. Using daily data from assets that compose seven market indexes around the world between 2008 and 2019, we tested neural networks with different settings of hidden layers and dropout rate, comparing various classification metrics, profitability and transaction costs levels to yield economic gain. The results indicated that the out-of-sample accuracy rate of the prediction converged to two values – besides the 50% value that represents the market efficiency, a "strange attractor" of 65% also was achieved consistently. On the other hand, the profitability of the strategies did not manage to significantly outperform the Buy-and-Hold strategy, even showing fairly large negative values for some hyperparameter combinations.

5.1 Introduction

Financial variables are hard to predict; over the decades, many scholars and market practitioners found many empirical evidences and stylized facts concerning the unpredictability of financial variables, ranging to stock prices to exchange rates. The efficient market hypothesis – which states that no economic agent can consistently yield higher returns than the market – remains as one of the most important theoretical results in finance. However, on the other side, numerous studies had been trying to analyze potential market inefficiencies and to predict the future trends of financial variables, such that the forecasting of the price and/or the directional movement of a stock price is still a largely debated and studied topic in finance.

Concerning this line of research, a wide variety of models have been tested and retested, and an equally wide variety of variables have been listed as potential sources

of useful information to accurately make those predictions. However, while the number of reported significant variables increases, the models become increasingly complex and harder to interpret in an intuitive and economically consistent way. In the philosophy of science, the "Occam's Razor" principle states that simpler solutions are superior to more complex ones, as parsimony itself is a desirable feature. The same principle appears in computer science and data analysis as "garbage in, garbage out" – that is, if a researcher feeds an analytical tool with bad data and/or useless information, the output is expected to be just as bad.

In terms of the generalization power of a model, the same reasoning can be applied: a good predictive model is one that describes well enough the observed data and is simple enough – that is, has a small level of complexity or noise. Therefore, the addition of a large amount of non-informative explanatory variables, even when able to provide a smaller in-sample error, boosts the level of noise in the model, making it actually worse for forecasting using future observations, as expressed in Hoeffding's inequality (equation **2.1**).

Besides providing a probabilistic upper bound for the generalization error E_{out} of a model's decision function, Hoeffding's inequality formalizes the trade-off between capacity and complexity for the construction of a good algorithm for generalizations. A good model for this purpose is one that lies on the optimal middle ground between describing well the data taken from the sample (a small in-sample error E_{in}), as well as deriving patterns for future and yet unseen data that are not way too complex, since the past data is filled not only with useful information but also an intrinsic component of noise, such that merely "memorizing" the past data tend to be not enough to cover satisfactorily future predictions. In statistics, this trade-off is also known as the "bias-variance dilemma".

The statistical challenges of high dimensionality and noisy independent variables are also addressed in Fan and Li (2006), which categorised feature selection problems under the "penalized likelihood" framework, which in turn expresses the main idea of Hoeffding's inequality: a good set of features is one that has good overall fitness and limited complexity, and adding non-informative variables would boost the latter in a larger magnitude than the former. Furthermore, Fan and Lv (2010) point out the computational setbacks when dealing with high dimensional data, in special non-concave penalized likelihood functions like the ones seen in neural networks – the backpropagation algorithm, for instance, converge to a local minimum for the loss function, and to actually find the global minimum can be very costly.

Specifically in finance, the advancements in this knowledge field have lead to an "overflow" of potential informative independent variables, as studies from the scientific literature use more and more different variables to explain a phenomenon or to predict a certain financial variable. In the scope of asset pricing, a big number of different factors

were proposed by the recent literature, leading to a proliferation of candidate factors, dubbed as a "factor zoo" by Cochrane (2011). Similarly, there is also a large number of variables listed by the literature for the stock price prediction, for indicators of both fundamentalist and technical analysis. Moreover, an also big number of technical analysis indicators are used by investors and market practitioners but not considered by the recent scientific literature in finance, making the variety of technical analysis indicators used for predictive means a "factor zoo" of its own. Therefore, this work intends to identify the components of said "zoo" and test the predictive performance of the features that compose it in light of machine learning techniques, as well as analyzing potential improvements of feature selection methods in selecting the most informative indicators and discarding the noisy or redundant ones.

5.2 Theoretical background

5.2.1 Factor zoo and feature selection in finance

Concerning the "factor zoo" in asset pricing models, Harvey, Liu and Zhu (2016) discussed the evolution of proposed models for the asset pricing problem, which evolved from a small number of factors to an extensive set of variables, listing 316 factors used by the specialized literature to model the cross-section of expected financial returns. However, the authors argued that most factors actually do not bring significant improvements on the models' performance, evidencing the presence of many non-informative and redundant variables that add more noise than explanatory power into the models. Furthermore, testing for the factors' significance using t-tests, Harvey, Liu and Zhu (2016) concluded that the high number of non-significant factors may likely indicate that a high number of research findings reported in financial economics papers actually do not hold, favoring instead classic and more parsimonious models.

Analyzing the effects of high dimensionality in finance, Kozak, Nagel and Santosh (2017) tested the effects of introducing nonlinear interactions between 130 factors up to degree-3 polynomials, subsequently applying dimensionality reduction techniques considering ℓ_1 and ℓ_2 regularizations to increase the model's sparsity. The results showed that a very small number of principal components are able to capture almost all of the out-of-sample explanatory power, while most principal components are non-informative; moreover, the introduction of additional regularized principal components does not hinder the model's sparsity, but does not improve predictive performance either.

Similar results were found in Hwang and Rubesam (2018) tested linear models from a set of 83 factors from the asset pricing literature "factor zoo", using a Bayesian estimation for seemingly unrelated regression models to search the best models. The authors tested for United States stocks from 1980 to 2016 and found out that only 10

factors were selected by the proposed method in some analyzed period, with only 5 to 6 showing actual significance on explaining the assets returns, although some selected factors in certain periods are not included in traditional factor models like Fama and French (1992), Fama and French (1996) and Fama and French (2015). Additionally, the study showed that the only factor that was consistently selected throughout the periods was the excess market return, which is consistent with the findings of Laloux et al. (1999), Nobi et al. (2013) and Sensoy, Yuksel and Erturk (2013), who found out that the market systematic risk is responsible for the largest eigenvalue of financial covariance matrices in many different financial markets and time periods, and that this top eigenvalue is significantly larger than the remaining ones, and that the vast majority of eigenvalues accounts for noisy information, falling into the theoretical bounds of a purely random Wishart covariance matrix.

Feng, Giglio and Xiu (2017) proposed a Two-Pass Regression approach with Double Selection LASSO with Monte Carlo simulations, incorporating the existence of model-selection mistakes that lead to an omitted variable bias to the models' coefficients estimation. Using 99 risk factors as inputs and testing for data between 1980 and 2016 for companies listed in NYSE, AMEX and NASDAQ with positive book equity, the authors reported that most recent proposed factors are statistically "redundant" or "useless", while relatively few were shown to be statistically "useful". Furthermore, the significance of the variables selected by the proposed method was shown to be more stable than the standard LASSO model.

Other applications of feature selection in financial contexts include Salcedo-Sanz et al. (2004), who proposed a feature selection method based on simulated annealing and Walsh analysis, coped with a classification Support Vector Machine, to predict insolvency for Spanish insurance companies. From a set of candidate variables with 21 accounting ratios, the proposed method generated subsets displaying the least noisy and redundant indicators. Similarly, Creamer (2012) proposed a trading algorithm for high frequency data of EURO STOXX 50 and DAX index Futures, applying machine learning techniques like boosting and bagging. The paper incorporated a big number of technical indicators, trading rules, and liquidity indicators, and the empirical analysis showed that whole set of variables, although containing indicators with a high degree of redundancy, yielded models' with smaller overall error rates. To predict the stock price direction of a Brazilian firm, Oliveira, Nobre and Zárate (2013) considered 46 variable distributed into macroeconomic variables, firm fundamentalist indexes, historical prices, and technical analysis indicators. The authors applied a filter feature selection method with correlation criterion, reducing the variable set to 18 features and applying them to a shallow ANN with one hidden layer, which yielded an out-of-sample accuracy rate of 87.5%.

5.2.2 Nonlinearity and machine learning in financial forecasting

Machine learning methods became a widely studied topic over the recent years due to their overall flexibility to the observed data, absence of restrictive assumptions of distribution and functional forms; instead, the basic premise is to "learn from the data" and identify potential non-intuitive patterns that contribute to better forecastings. One of the main features of machine learning methods is the insertion of nonlinearities, allowing the modeling of high degrees of nonlinearity in a reduced number of functions and hyperparameters. For instance, a neural network is essentially a linear regression with "chunks" of nonlinearity; a classification support vector machine is a linear separator in a feature space with an arbitrarily high dimension, depending on the Kernel function applied.

Mullainathan and Spiess (2017) discussed the connections between machine learning methods and econometrics, showing that both generalized linear methods – like LASSO and ridge regression – and machine learning algorithms like random forest, K-nearest neighbors and neural networks can be represented as a particular case of minimizing the in-sample loss with a complexity restriction, analogous to the framework stated in Hoeffding's inequality. In this sense, the insertion of irrelevant variables as inputs for machine learning models can potentially be even more harmful to the model's predicting performance.

While the presence of nonlinear interaction variables can reveal additional patterns and joint significances between the original variables, it also augments the "zoo size" even more, making the presence of a non-informative feature more problematic due to the possible spreading of its noisy effect to other useful factors: for instance, taking a set of n variables, the number of cross-interactions of degree 2 of those variables would yield $n + \binom{n}{2}$ variables, and $n + n(n-1) + \binom{n}{3}$ variables more of degree 3, and so on. While the number of potentially useful regressors already leads to a high level of noise and high proneness of overfitting, taking the analysis to nonlinear relationships may hinder the model's overall predictive power even further.

In traditional econometrics, using highly correlated variables leads to the well-known problem of multicollinearity, which leads to larger standard errors on the estimates and interfering on the model's inference, making it less robust, as well as increasing the chances on inaccuracies on the numerical optimization of the computer algorithm. In machine learning models the effects are similar, as the model tends to overfit and make inaccurate predictions for new data. This can be seen in Guresen, Kayakutlu and Daim (2011), in which ANNs and ensemble models were tested to predict the daily NASDAQ Index using a sample of 182 days between 2008 and 2009. The authors tested a hybrid neural network that introduced new input variables constructed from GARCH and EGARCH models, and the results of these methods were shown to be worse than the

conventional multilayer perceptron, indicating that the added features introduced more noise than actual explanatory power.

Conversely, as discussed in Guyon and Elisseeff (2003), even highly correlated variables (positively or negatively) does not necessarily imply in the absence of variable complementarity, since there may be nonlinear dependencies that are actually informative. In this sense, since machine learning methods rely heavily on nonlinear interactions, an apparently redundant variable subset, although can make the model more prone to overfitting and making inaccurate predictions, can actually yield good predictions that when applied to a linear model.

When dealing with a high number of variables, a common approach to extract the most relevant information out of the feature set is to apply Principal Component Analysis (PCA) to take linear combinations that account for the largest proportions of explained variance. However, while PCA is a practical way, the principal components do not have an immediate implication in real life, unlike feature selection methods which take variables "as a whole", maintaining their respective economic and financial interpretations. Furthermore, PCA forces the principal components to be orthogonal – that is, linearly non-correlated. Thus, when nonlinear relationships of the candidate variables are being accounted, the original interpretability suffers an additional loss, for it may be unnatural to imagine the intuition of a polynomial or exponential version of a variable with clear economic meaning. In this sense, in this paper we opted to take subsets of individual variables instead of combinations of features, given that the principal components themselves are already hard to interpret, and using them as inputs to a nonlinear predictive model like ANN would further hinder the model's interpretability.

Moghaddam, Moghaddam and Esfandyari (2016) applied ANN to forecast the daily NASDAQ stock exchange rate with data from January 2015 to June 2015 using past prices and the day of the week as input variables. Fenghua et al. (2014) applied a SVM combined with singular spectrum analysis to predict the daily closing price of the SSE Composite Index from 2009 to 2013. Nayak, Pai and Pai (2016) applied Boosted Decision Tree, Logistic Regression and SVM to predict the Indian stock market trend using historic prices and market sentiment measured by posts on Twitter.

Qiu, Song and Akagi (2016) emphasized the theoretical capacity of neural networks to approximate any nonlinear continuous function, reiterating that the nonlinear behavior of financial series is one of the main challenges of forecasting stock market returns. ANNs were applied to predict the return of the Nikkei 225 Index using macroeconomic variables like monetary base, interest rates, trade flow and industrial production as inputs. Genetic algorithms and simulated annealing were combined with neural networks to improve the prediction accuracy and to overcome the local convergence problem of the backpropagation algorithm.

The improvements of nonlinearity in financial applications are also discussed in Gu, Kelly and Xiu (2018), who applied various machine learning methods – namely principal components regression, partial least squares, generalized linear models, boosted regression trees, random forests and artificial neural networks – and compared them to simple and penalized (ridge regression, LASSO and elastic net) linear models to measure the risk premium of financial assets using data of nearly 30000 financial assets from NYSE and NASDAQ between 1957 and 2016. The results presented empirical evidences favoring machine learning models in terms of providing a more accurate description of the price oscillation patterns of the analyzed assets in comparison to traditional statistical methods; the authors credited the predictive performance gain to the introduction of nonlinear predictor interactions from those methods, which are not considered by commonly used econometric approaches. Moreover, the authors reported that all models converged to a similar set of "essential predictors" composed mainly by variations in the assets' liquidity and volatility.

5.2.3 Technical analysis indicators and machine learning in stock price predictions

Technical analysis indicators are commonly used tools in investment evaluation, financial trading and portfolio selection. As detailed in table **23**, there is a lengthy list of technical indicators used in the literature as variables to predict a stock's price movement; however, at the same time, many indicators are highly correlated, with some of them actually being defined by simple mathematical operations or combinations between other technical indicators which are also candidate criteria for the investor's decision. For example, the Stochastic D% indicator is simply an arithmetic mean of the Stochastic K% indicator for the last n periods, with Stochastic D% itself calculated using the maximum and minimum prices over the last n periods, which are themselves used as separate indicators; both the Momentum (MOM) and the Rate of Change (ROC) describe the variation of the closing price relative from n periods before, and the only difference between them is that the former states the difference in absolute variation, while the later gives away the percentage variation.

Given the high degree of "overlapping" between technical analysis indicators, a previous feature selection filtering which variables are most relevant to predict a stock movement is an important issue to avoid prediction problems arising from redundant variables like multicollinearity and overfitting. Therefore, this paper aimed to observe the improvements in out-of-sample prediction application of feature selection methods for the prediction of the stock price direction. Besides, we evaluated the effect of feature selection in machine learning models, providing insights about the best combination of machine learning model and feature selection method, as well as the emergence of chaotic

behavior of machine learning methods.

The application of technical analysis on stock markets was analyzed in Nazário et al. (2017)'s literature review, in which 85 papers published between 1959 and 2016 were classified in terms of the chosen markets, methodologies, consideration of risk and transaction costs, operational tools, among other categories. The analysis showed that artificial neural networks are a widely used tool in this literature, mainly due to their consistency for small-range data; in special, the popularity of this technique increased in the last years covered by the analysis, which coincided with the overall interest in machine learning applications in finance.

Given the popularity and empirical effectiveness of machine learning models in financial forecasting, this class of models has been actively being applied to the prediction of the financial stock prices. As shown in an extensive review by Henrique, Sobreiro and Kimura (2019), who mapped 57 papers published in high-impact journals, the application of machine learning techniques to the prediction of financial stock prices is still a highly debated research topic in the recent literature, both regarding the forecasting of the market movement direction (a classification problem) and the magnitude of the movement itself (a regression problem). More specifically, the paper indicated that the input variables considered by the models are basically divided into fundamentalist and technical indicators, with the most prominent machine learning methods applied to financial market predictions being Artificial Neural Networks and Support Vector Machines, as well as their respective extensions. Therefore, we tested different setting of ANN models to verify the predictive ability of the models and to analyze the effects of the pre-application of feature selection methods.

Henrique, Sobreiro and Kimura (2018) used five technical analysis indicators to predict stock price from Brazilian, American and Chinese financial markets using Support Vector Regression (SVR) for high frequency data, using the Random Walk as the benchmark. The paper argues in favor of the predictive power of the SVR models, especially in periods of lower market volatility.

Costa et al. (2015) tested the performance of trading strategies in comparison to the buy-and-hold strategy based on technical analysis indicators for 198 stocks of the Brazilian market, and analyzed their respective predictability for the market trends under various circumstances of transaction costs. The paper reports that the proposed strategies obtained returns larger than the invested value, but have reduced predictive power for the stocks' future prices.

Żbikowski (2015) applied a volume weighted extension of a classification SVM to predict the stock price direction for 20 US stocks between 2003 and 2013 using 7 technical analysis indicators. The paper also tested a feature selection method based on Fisher Score ranking, and although the application of this procedure has reached a better performance

of trading strategies, the author advised caution and recommended additional researches regarding the effectiveness of other feature selection methods. Specifically, in this paper we further investigated the effects of feature selection before applying a machine learning technique for a much larger initial set of indicators which included all 7 indicators used in Żbikowski (2015).

Zhu et al. (2015) performed a similar analysis for two Chinese stock exchange indexes between 1991 and 2013. Comparing trading range break, fixed-length moving average and variable moving average rules, the study's empirical results showed that the former outperformed the others in terms of profitability; specifically, short-term moving average rules worked better than long-term ones. Moreover, White's Reality Check test indicated that the best trading signals from variable moving average and trading range break outperformed the buy-and-hold strategy in a scenario without transactions costs; when they are taken into account, however, there were no statistical evidences towards the superiority of the technical analysis trading rules.

Alhashel, Almudhaf and Hansz (2018) tested 22 technical analysis trading rules for indexes from 1995 to 2015 of nine financial markets in Asia, namely China, Hong Kong, Indonesia, Japan, Malaysia, Philippines, Singapore, Taiwan, and Thailand. Controlling for transaction costs and strategy risk, the results found evidences of market inefficiency in four of the analyzed markets, implying the existence of predictive power of technical analysis indicators on those markets; on the other five markets, though, the profitability did not outperform the market gains.

Nakano, Takahashi and Takahashi (2018) proposed trading strategies using intra-day bitcoin price data applying ANNs for the prediction of the returns. Technical analysis indicators and historical return data were used as input variables, and both shallow and deep network architectures were tested. Even considering the effect of transaction costs, the risk-adjusted profitability of the proposed method was reported to be significantly greater than the buy-and-hold strategy, especially during a period when bitcoin prices suffered a large drawback.

Weng, Ahmed and Megahed (2017) combined online data information collected from "knowledge bases" Google and Wikipedia with traditional time-series and financial technical analysis indicators to build a trading expert system that operates on a daily periodicity. Machine learning techniques (decision trees, ANN and SVM) were used as the predictive tool of the proposed system. Even though the sample consisted of only one company, the paper reported an 85% directional accuracy for the predictions and claimed improvement over the results of similar works in the literature.

Patel et al. (2015b) proposed a two-stage fusion model between ANN, Random Forest and Support Vector Regression (SVR) to predict the value of two Indian stock market indexes (NIFTY 50 and BSE SENSEX), with the first stage estimating the pa-

rameters used in the second stage. Using data from 2003 to 2012 and 10 technical analysis indicators as independent variables, the authors reported that the two-stage procedure led to a diminishment of the overall out-of-sample prediction error levels in comparison to single step versions of the adopted machine learning models. Among the machine learning models applied individually, ANN and SVR exhibited superior overall performance than Random Forest, while the ANN-SVR hybrid yielded the lowest error metrics between the tested combinations.

The "factor zoo" of technical analysis indicators is displayed in table **23**: in this table are listed technical analysis indicators used as independent variables in papers published in high impact journals over the last 20 years (1999-2018) that applied machine learning models in the forecasting of the value or the direction of stock prices of financial market indexes. On the "trader side", four specialized websites that offer financial services and technical analysis softwares (Fidelity Investments, Trading Technologies, StockCharts, and TradingView) were analyzed and indicators documented in their respective websites that were not used in any of the academic researchers were listed in table **24**.

Variable	References
Simple Moving Average	Thawornwong, Enke and Dagli (2003), Chang and Fan (2008), Chang et al. (2009), Huang and Tsai (2009) Yu et al. (2009), Vanstone and Finnie (2010), Kara, Boyacioglu and Baykan (2011), Chang et al. (2012) Creamer (2012), Oliveira, Nobre and Zárate (2013), Chen, Cheng and Tsai (2014), Patel et al. (2015a) Patel et al. (2015b), Chiang et al. (2016), Novak and Velušček (2016), Weng, Ahmed and Megahed (2017) Alhashel, Almudhaf and Hansz (2018), Henrique, Sobreiro and Kimura (2018)
Weighted Moving Average	Kara, Boyacioglu and Baykan (2011), Patel et al. (2015a), Patel et al. (2015b), Novak and Velušček (2016) Alhashel, Almudhaf and Hansz (2018), Henrique, Sobreiro and Kimura (2018)
Exponential Moving Average	Tay and Cao (2001), Ang and Quek (2006), Yu et al. (2009), Vanstone and Finnie (2010) Creamer (2012), Ticknor (2013), Novak and Velušček (2016), Chen et al. (2017) Weng, Ahmed and Megahed (2017), Nakano, Takahashi and Takahashi (2018), Alhashel, Almudhaf and Hansz (2018)
Momentum	Kim and Han (2000), Kim (2003), Yu et al. (2009), Kara, Boyacioglu and Baykan (2011), Creamer (2012) Oliveira, Nobre and Zárate (2013), Chen, Cheng and Tsai (2014), Patel et al. (2015a), Patel et al. (2015b) Chiang et al. (2016), Novak and Velušček (2016), Weng, Ahmed and Megahed (2017)
Stochastic K%	Kim and Han (2000), Kim (2003), Thawornwong, Enke and Dagli (2003), Kwon and Moon (2007), Chang et al. (2009) Huang and Tsai (2009), Yu et al. (2009), Vanstone and Finnie (2010), Kara, Boyacioglu and Baykan (2011) Chang et al. (2012), Ticknor (2013), Oliveira, Nobre and Zárate (2013), Chen et al. (2014) Patel et al. (2015a), Patel et al. (2015b) Nakano, Takahashi and Takahashi (2018), Alhashel, Almudhaf and Hansz (2018)
Stochastic D%	Kim and Han (2000), Kim (2003), Kwon and Moon (2007), Chang and Fan (2008) Chang et al. (2009), Huang and Tsai (2009), Yu et al. (2009), Kara, Boyacioglu and Baykan (2011) Chang et al. (2012), Ticknor (2013), Oliveira, Nobre and Zárate (2013), Chen et al. (2014) Patel et al. (2015a), Patel et al. (2015b), Nakano, Takahashi and Takahashi (2018)
Slow Stochastic D%	Kim and Han (2000), Kim (2003), Yu et al. (2009)
Relative Strength Index	Kim and Han (2000), Kim (2003), Thawornwong, Enke and Dagli (2003), Armano, Marchesi and Murru (2005) Kwon and Moon (2007), Chang and Fan (2008), Chang et al. (2009), Huang and Tsai (2009) Yu et al. (2009), Vanstone and Finnie (2010), Kara, Boyacioglu and Baykan (2011), Rodríguez-González et al. (2011) Chang et al. (2012), Creamer (2012), Ticknor (2013), Oliveira, Nobre and Zárate (2013) Chen, Cheng and Tsai (2014), Patel et al. (2015a), Patel et al. (2015b), Novak and Velušček (2016) Weng, Ahmed and Megahed (2017), Nakano, Takahashi and Takahashi (2018), Alhashel, Almudhaf and Hansz (2018) Henrique, Sobreiro and Kimura (2018)
Moving Average Convergence-Divergence	Tay and Cao (2001), Thawornwong, Enke and Dagli (2003), Armano, Marchesi and Murru (2005), Kwon and Moon (2007) Chang and Fan (2008), Chang et al. (2009), Huang and Tsai (2009), Yu et al. (2009) Vanstone and Finnie (2010), Kara, Boyacioglu and Baykan (2011), Chang et al. (2012), Creamer (2012) Oliveira, Nobre and Zárate (2013), Chen, Cheng and Tsai (2014), Chen et al. (2014), Patel et al. (2015a) Patel et al. (2015b), Chiang et al. (2016), Nakano, Takahashi and Takahashi (2018), Alhashel, Almudhaf and Hansz (2018)
William's R%	Kim and Han (2000), Kim (2003), Chang et al. (2009), Huang and Tsai (2009)

Accumulation/Distribution Oscillator	Kara, Boyacioglu and Baykan (2011), Chang et al. (2012), Ticknor (2013), Oliveira, Nobre and Zárate (2013) Chen, Cheng and Tsai (2014), Patel et al. (2015a), Patel et al. (2015b), Alhashel, Almudhaf and Hansz (2018) Kim and Han (2000), Kim (2003), Yu et al. (2009), Kara, Boyacioglu and Baykan (2011) Patel et al. (2015a), Patel et al. (2015b), Alhashel, Almudhaf and Hansz (2018), Henrique, Sobreiro and Kimura (2018)
Commodity Channel Index	Kim and Han (2000), Kim (2003), Yu et al. (2009), Kara, Boyacioglu and Baykan (2011) Patel et al. (2015a), Patel et al. (2015b), Novak and Velušček (2016), Alhashel, Almudhaf and Hansz (2018)
Rate of Change	Kim and Han (2000), Tay and Cao (2001), Kim (2003), Armano, Marchesi and Murru (2005) Chang et al. (2009), Yu et al. (2009), Creamer (2012), Novak and Velušček (2016), Weng, Ahmed and Megahed (2017) Alhashel, Almudhaf and Hansz (2018)
Disparity	Kim and Han (2000), Kim (2003), Yu et al. (2009), Weng, Ahmed and Megahed (2017)
Price Oscillator	Kim and Han (2000), Kim (2003), Yu et al. (2009)
Psychological line	Chang and Fan (2008), Huang and Tsai (2009), Chen, Cheng and Tsai (2014)
Directional Indicator Up	Huang and Tsai (2009)
Directional Indicator Down	Huang and Tsai (2009)
Bias	Chang and Fan (2008), Chang et al. (2009), Huang and Tsai (2009), Chen, Cheng and Tsai (2014), Chang et al. (2012)
Volume Ratio	Huang and Tsai (2009)
A Ratio	Huang and Tsai (2009)
B Ratio	Huang and Tsai (2009)
Average True Range	Vanstone and Finnie (2010), Henrique, Sobreiro and Kimura (2018)
Bollinger Band Upper	Creamer (2012), Oliveira, Nobre and Zárate (2013), Alhashel, Almudhaf and Hansz (2018)
Bollinger Band Lower	Creamer (2012), Oliveira, Nobre and Zárate (2013), Alhashel, Almudhaf and Hansz (2018)
Directional Movement Indicator	Alhashel, Almudhaf and Hansz (2018)
Keltner Channel Upper Band	Alhashel, Almudhaf and Hansz (2018)
Keltner Channel Lower Band	Alhashel, Almudhaf and Hansz (2018)
Triangular Moving Average	Alhashel, Almudhaf and Hansz (2018)
Moving Average Envelope Upper	Alhashel, Almudhaf and Hansz (2018)
Moving Average Envelope Lower	Alhashel, Almudhaf and Hansz (2018)
Rex Oscillator	Alhashel, Almudhaf and Hansz (2018)
Negative Volume Index	Creamer (2012)
Positive Volume Index	Creamer (2012)
Volume Adjusted Moving Average	Chavarnakul and Enke (2009)
Highest Price ratio	Vanstone and Finnie (2010)
Lowest Price ratio	Vanstone and Finnie (2010)
Opening price	Oliveira, Nobre and Zárate (2013)
Closing price	Oliveira, Nobre and Zárate (2013)
Minimum price	Oliveira, Nobre and Zárate (2013)
Minimum price	Oliveira, Nobre and Zárate (2013)

Volume	Chang et al. (2009), Oliveira, Nobre and Zárate (2013), Novak and Velušček (2016)
Volume Momentum	Chiang et al. (2016)
Moving Price Level Percentage	Chiang et al. (2016)
Percent Price Oscillator	Chiang et al. (2016)
Parabolic Stop and Reverse	Novak and Velušček (2016), Alhashel, Almudhaf and Hansz (2018)
On Balance Volume	Tay and Cao (2001), Creamer (2012), Oliveira, Nobre and Zárate (2013), Nakano, Takahashi and Takahashi (2018)
Average Directional Movement Index	Vanstone and Finnie (2010)
Volatility	Tay and Cao (2001)
Money Flow Index	Thawornwong, Enke and Dagli (2003)
Variance Ratio	Yu et al. (2009)
Linear Regression Slope	Yu et al. (2009)

Table 23 – Technical analysis indicators used in recent financial prediction studies that applied machine learning models

Variable	References
Acceleration Band Up	Trading Technologies (2019)
Acceleration Band Down	Trading Technologies (2019)
Accumulation/Distribution Index	Trading Technologies (2019), Fidelity Investments (2019)
Money Flow Multiplier	StockCharts (2019)
Accumulation Distribution Line	StockCharts (2019), TradingView (2019)
Absolute Price Oscillator	Trading Technologies (2019), Fidelity Investments (2019)
Aroon Indicator Positive	Trading Technologies (2019), StockCharts (2019) Fidelity Investments (2019), TradingView (2019)
Aroon Indicator Negative	Trading Technologies (2019), StockCharts (2019) Fidelity Investments (2019), TradingView (2019)
Aroon Oscillator	Trading Technologies (2019), StockCharts (2019)
Average True Range Percent	Fidelity Investments (2019)
Average Volume	Fidelity Investments (2019)
Bollinger Band Width	StockCharts (2019), Fidelity Investments (2019) TradingView (2019)
Bollinger Band %B	StockCharts (2019), Fidelity Investments (2019) TradingView (2019)
Band Width	Trading Technologies (2019)
Chaikin Money Flow	StockCharts (2019), Fidelity Investments (2019) TradingView (2019)
Chaikin Oscillator	StockCharts (2019), TradingView (2019)
Chaikin Volatility	Fidelity Investments (2019)
Chande Momentum Oscillator	Trading Technologies (2019), Fidelity Investments (2019)
Chandelier Exit Long	StockCharts (2019)
Chandelier Exit Short	StockCharts (2019)
Choppiness Index	TradingView (2019)
Coppock Curve	StockCharts (2019)
Detrended Price Oscillator	StockCharts (2019), Fidelity Investments (2019) TradingView (2019)
Donchian channel	Cavendish Astrophysics (2011), TradingView (2019)
Double Exponential Moving Average	Trading Technologies (2019)
Double Smoothed Stochastic	Fidelity Investments (2019)
Ease of Movement	StockCharts (2019), TradingView (2019)
Force Index	StockCharts (2019), TradingView (2019)
Hull Moving Average	Fidelity Investments (2019)
Kaufman's Adaptive Moving Average	StockCharts (2019)

Linear Regression Intercept	Trading Technologies (2019), StockCharts (2019) Fidelity Investments (2019)
MACD Histogram	StockCharts (2019)
Mass Index	StockCharts (2019)
Raw Money Flow	Fidelity Investments (2019)
Midpoint	Trading Technologies (2019)
Midprice	Trading Technologies (2019)
Normalized Average True Range	Trading Technologies (2019)
Typical Price	Fidelity Investments (2019)
Standard Support 1	Fidelity Investments (2019)
Standard Support 2	Fidelity Investments (2019)
Standard Resistance 1	Fidelity Investments (2019)
Standard Resistance 2	Fidelity Investments (2019)
Fibonacci Support 1	Fidelity Investments (2019)
Fibonacci Support 2	Fidelity Investments (2019)
Fibonacci Resistance 1	Fidelity Investments (2019)
Fibonacci Resistance 2	Fidelity Investments (2019)
Demark Pivot Point	Fidelity Investments (2019)
Demark Support	Fidelity Investments (2019)
Demark Resistance	Fidelity Investments (2019)
Price Channel Upper	Trading Technologies (2019), StockCharts (2019)
Price Channel Lower	Trading Technologies (2019), StockCharts (2019)
PPO Histogram	StockCharts (2019)
Percentage Volume Oscillator	StockCharts (2019)
PVO Histogram	StockCharts (2019)
Price Volume Trend	Trading Technologies (2019), TradingView (2019)
Pring's Know Sure Thing Oscillator	StockCharts (2019), TradingView (2019)
Pring's Special K	StockCharts (2019)
Relative Vigor Index	Fidelity Investments (2019)
Standard Error	Fidelity Investments (2019)
Stochastic RSI	Fidelity Investments (2019), TradingView (2019)
Triple Exponential Moving Average	Trading Technologies (2019)
Triple Exponential Moving Average Oscillator	Trading Technologies (2019), StockCharts (2019) TradingView (2019)
True Strength Index	StockCharts (2019)
Typical Price	Fidelity Investments (2019)
Ulcer Index	StockCharts (2019)
Ultimate Oscillator	Trading Technologies (2019), StockCharts (2019)

	Fidelity Investments (2019), TradingView (2019)
Volume Oscillator	Fidelity Investments (2019)
Volume Price Trend	Cavendish Astrophysics (2011)
Volume Weighted Average Price	Trading Technologies (2019), StockCharts (2019) TradingView (2019)
Vortex Indicator Positive	StockCharts (2019)
Vortex Indicator Negative	StockCharts (2019)
Welles Wilder's Smoothing Average	Trading Technologies (2019)

Table 24 – Technical analysis indicators not used in recent financial studies

5.3 Method

5.3.1 Logistic Regression and Artificial Neural Networks

The most well-known classification algorithm is the logistic regression (also known as the "logit model"), which is basically a linear regression model for the log-odds of the probability of success of a Bernoulli experiment $p(\boldsymbol{x})$ conditioned to a vector of observed independent variables $\boldsymbol{x}_{(kx1)}$, such that:

$$log\left(\frac{p_i(\boldsymbol{x})}{1-p_i(\boldsymbol{x})}\right) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i} \tag{5.1}$$

which can be rewritten as:

$$p_i(\boldsymbol{x}) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_{1,i}+\cdots+\beta_k x_{k,i})}} = \frac{e^{\beta_0+\beta_1 x_{1,i}+\cdots+\beta_k x_{k,i}}}{1+e^{\beta_0+\beta_1 x_{1,i}+\cdots+\beta_k x_{k,i}}} \tag{5.2}$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$ is known as the sigmoid function (or standard logistic function)

In summary, the logistic regression is a linear regression whose output is "squashed" into the range $[0, 1]$ through the sigmoid function. Given that the log-odd can be interpreted as the ratio between the probability of success and the probability of failure of the Bernoulli experiment, usually the cutoff $p_i(\boldsymbol{x}) = 0.5$ is used as a classifying rule for binary dependent variables, with the prediction being "class 1" if $p_i(\boldsymbol{x}) > 0.5$ and "class 0" if otherwise.

While being simple and providing an easy interpretation, the logistic regression has, by construction, an assumption of linearity, which makes a major limitation of this model. Over recent years, many other nonparametric models showed better empirical performance in classification tasks, hence becoming increasingly popular among researchers due to their flexibility, notably in financial contexts: as discussed in Hsu et al. (2016), machine learning methods were shown to consistently outperform traditional econometrics models in a various range of applications, including problems in finance such as stock market prediction, portfolio analysis, asset pricing, and risk management. One of the most used machine learning methods are Artificial Neural Networks (ANN) as well as their many extensions, such as deep networks, recurrent networks, convolutional networks, among others. Their applications range from image processing, text translating to chromosome mapping and financial forecasting; specifically in finance, references that used ANNs were briefly summarized in section **5.2**.

While the big variety of ANN extensions differ in functional forms and complexities, in essence, an ANN is but a recursive application of linear models and "chunks" of nonlinearity. Algebraically, while a linear regression model can be expressed as $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w}$,

with $\boldsymbol{Y}$ being a vector of dependent variables, $\boldsymbol{X}$ a matrix of observed independent variables and $\boldsymbol{w}$ a vector of parameters, an ANN can be generally expressed as:

$$\boldsymbol{y} = \boldsymbol{\psi_\ell}(...\boldsymbol{\psi_2}(\boldsymbol{\psi_1}(\boldsymbol{X}\boldsymbol{W_1})\boldsymbol{W_2})...\boldsymbol{W_\ell})\boldsymbol{w} \tag{5.3}$$

where $\psi(.)$ are activation functions – where the nonlinearity is introduced – and ℓ is the number of hidden layers.

In other words, an ANN is simply a sequence of linear regressions, and the linear regression itself is actually an ANN with one hidden layer, with $\ell = 1$ and $\psi(\boldsymbol{x}) = \boldsymbol{x}$. Similarly, the previously discussed logistic regression is also an ANN with $\ell = 1$ and $\psi(\boldsymbol{x})$ equal to the logistic function $\sigma(\boldsymbol{x}) = \frac{1}{1+e^{-\boldsymbol{x}}}$, which is one of the most popular activation functions in ANN applications.

Instead of using only one hidden layer, an ANN can be specified with an arbitrary degree of "deepness" by stacking more layers, which allows the algorithm to learn more abstract knowledge representations. For instance, in image recognition, the first layers focus on simpler tasks like identifying the contrast of the pixels to preliminary define the contours of the objects in the image; as the inputs go deeper into the layers, more complex patterns like edges and lines are learned, and in the deepest layers the neurons specialize in identifying actual objects, such as eyes and ears (GOODFELLOW et al., 2016). In financial applications, the same reasoning is valid, as discussed by Heaton, Polson and Witte (2016): by using an ANN with more hidden layers, the algorithm becomes able to learn stylized facts from financial data such as volatility clustering and the leverage effect; in factor models, an ANN generalizes cross-interactions between the factor as a hierarchical nonlinear factor model. Regarding technical analysis indicators, it is expected for ANNs to learn the trading rules that derive from them – for instance, a buy or sell signal arising from the crossing of a short-term and a long-term moving average, as summarized in the MACD indicator. Moreover, a deep ANN can theoretically provide a joint analysis for different indicators that may give away different actions for investors – for example, if some indicators give a buy signal while others give a sell signal, a deep ANN is able to consider more abstract market dynamics that each indicator is modeling and their joint effect on the investor's ultimate decision.

Nonetheless, a deeper network structure, while allowing the algorithm to learn more complex structures, also makes it more prone to overfitting – that is, the ANN may simply "memorize" the in-sample data alongside with the noisy information specific to those observations, making it worse for generalizations. When dealing with financial data, this issue is particularly relevant, as pointed out in Heaton, Polson and Witte (2017). In this sense, in the literature of financial applications, there is no consensus concerning the effects of introducing additional hidden layers in ANNs on the model's

out-of-sample predictive performance. For example, in Nakano, Takahashi and Takahashi (2018)'s experiments on the profitability of trading strategies for bitcoin, deeper ANNs yielded better results, indicating that more levels of interaction between the variables can help to reveal more complex patterns in the data. On the other hand, in Gu, Kelly and Xiu (2018)'s application on risk premia of US Stocks assets, neural networks with different numbers of hidden layers were tested, and the deep architecture neural networks showed worse results than shallower networks, evidencing also the possibility of overfitting when dealing with data with high levels of noise. Therefore, in this paper we tested for ANNs with different numbers of hidden layers and analyzed the effects of previously applying feature selection methods in the original set of technical analysis indicators.

5.3.2 Regularization and Dropout

Dropout, introduced by Srivastava et al. (2014), is a regularization method commonly used in neural network applications to avoid overfitting in the training process. In each iteration, instead of computing every possible parameter throughout the network in the backpropagation, each neuron is activated with a probability $1 - p$, with p being the dropout rate. The motivation of doing so is to force the neural network to learn from a broader range of "paths" instead of attributing higher weights to interactions between specific neurons, consequently ignoring other possible interactions. In this sense, dropout can be regarded as an indirect form of feature selection for nonlinear interactions of the original variables, making the ANN converge to a set of weights that are larger for important interactions and closer to zero for irrelevant ones.

In each training epoch, the dropout randomly assigns a percentage of features as zero, such that no further interactions will be derived from the zeroed neurons. The mechanism avoids the excessive exploitation of a previously neuron path, thus decreasing the chance of memorizing a path that conveniently yielded a smaller rate of error. When one neuron of this path is removed, the network will be forced to consider other potential paths that minimize the in-sample error, thus allowing it to learn other potentially useful patterns for predicting using out-of-sample data.

While dropout is already a well-known procedure in recent machine learning studies, its use as a regularization tool is still scarce in the financial applications: while this technique was employed in Heaton, Polson and Witte (2017) for the construction of "deep portfolios", specifically concerning applications that used technical analysis indicators and ANNs to stock prices prediction, dropout was not applied in any of the papers analyzed in review papers Nazário et al. (2017) and Henrique, Sobreiro and Kimura (2019). Bearing in mind the large number of features present in our application, an additional mechanism to control overfitting and the ANNs'complexity is desirable in terms of predicting effectiveness. Therefore, in this paper we applied this method and verified whether this

technique managed to further improve the predictive performance of both the original feature set and the refined set after applying feature selection.

5.3.3 Feature selection methods

The survey paper of Chandrashekar and Sahin (2014) classified feature selection methods in three broad categories: filter methods, in which the features are ranked according to their respective conditional dependencies to the class labels; wrapper methods, where an algorithm searches for the feature subset with the best predictive performance in a validation dataset; and embedded methods, in which the feature selection occurs simultaneously with the training process without splitting the dataset. Concerning the category of filter methods, Chandrashekar and Sahin (2014) point out that approaches like feature ranking by criteria like covariance and mutual information tend to ignore inter-feature dependencies, notably highly nonlinear ones; moreover, the authors report that there is no ideal method indicated by the literature regarding the choice of the size of the filtered feature set. As shown in the experiments of John, Kohavi and Pfleger (1994), the subsets yielded from wrapper selection models generated more parsimonious classification decision trees in comparison to filter selection models; additionally, in this paper we applied ANNs, a method whose main differentials are precisely the introduction of nonlinearities and the learning of abstract dependency structures between the input features. Hence, we did not consider filter feature selection methods and tested only wrapper and embedded methods instead. Additionally, Chandrashekar and Sahin (2014) subclassified wrapper methods in sequential selection algorithms and heuristic search algorithms. In this sense, we tested three feature selection methods, described in the following subsections:

5.3.3.1 Sequential Forward Floating Selection (SFFS)

The Sequential Forward Floating Selection (SFFS) algorithm, proposed by Pudil, Novovičová and Kittler (1994), is a sequential selection algorithm that combines a forward wrapper selection method (Sequential Feature Selection – SFS) with a backward one (Sequential Backward Selection – SBS). The SFS starts with an empty set of features and adds recursively the variable that gives away the highest improvement in the classification performance until the refined subset reaches an user-defined size parameter d. SBS works in a similar way, starting from the full set of all variables and removes features that give away the lowest decrease in prediction performance.

As both SFS and SBS are greedy approaches, variable subsets that present a high improvement when jointly considered can be missed by these algorithms. SFFS, on the other hand, adds more flexibility to the basic SFS by adding a step to exclude already included features instead of keeping them permanently. Thus, the SFFS alternates a forward step from SFS with a backward step in which previously added features are ex-

cluded while a new best feature subset is obtained through the exclusion. The superiority of SFFS over SFS is discussed in Reunanen (2003), who compared both frameworks for datasets with classification tasks from various knowledge fields collected from the UCI Machine Learning Repository.

Concerning stock direction prediction, Lee (2009) applied the SFFS algorithm to perform wrapper feature selection on a set of 29 variables composed by financial indexes, currency quotations, and commodities futures to predict the NASDAQ index direction; the selected feature subset was then fitted into a classification SVM and a standard ANN. Moreover, the study compared the effectiveness of other filter feature selection methods based on three criteria, namely information gain, symmetrical uncertainty and correlation, and SFFS was the algorithm that yielded the highest improvements over its full feature set counterpart for both models. The steps of the SFFS algorithm can be summarized as below:

```
<D> = 0
<performance_best> = 0
WHILE (D < K):
    EXECUTE <1-step SFS>
    KEEP <performance_forward>
    D = D + 1
    IF (performance_forward > performance_best)
        performance_best = performance_forward
    <performance_backward> = INFINITY
    WHILE (performance_backward > performance_best)
        EXECUTE <1-step SBS>
        D = D - 1
        KEEP <performance_backward>
            IF (performance_backward > performance_best)
                performance_best = performance_backward
            ELSE
                UNDO <1-step SBS>
                D = D + 1
END
```

5.3.3.2 Tournament screening (TS)

The tournament screening (TS) algorithm, proposed by Chen and Chen (2009), is a heuristic search algorithm which generates candidate features based on the best features of mutually exclusive subsets. The variables are recursively split into smaller groups and a "tournament" takes place inside each subset, which the features that "survive" the contests being classified as the best ones. Therefore, the main idea of the tournament screening is analogous from a genetic algorithm, in which the "strongest" offsprings crossover amongst themselves while the weakest are gradually eliminated (CHANDRASHEKAR; SAHIN, 2014).

In the TS algorithm, the set of original variables are randomly subdivided into disjoint subsets, and inside each subset, a verification model is fitted and the variable with the least contribution or statistical significance is excluded from the group. The remaining variables from all subsets are aggregated and attributed again to new mutually exclusive subsets, repeating the process of recursive elimination until the number of remaining is reduced to a user-specified number. As pointed out by Alves (2014) and Saavedra (2015), the tournament screening is a good alternative for parametric models of high dimensionality, in special when the number of features is so high to the point where the number of degrees of freedom is not sufficient for the joint estimation of all parameters specified in the model. In those cases, the application of tournament screening allows the parameters to be estimated within each subgroup, assuming that the null parameters will ultimately be estimated as non-influential values inside those subgroups.

The pseudocode for the TS algorithm is displayed below:

```
WHILE length(feature_set) > K:
    SPLIT <feature_set> into P mutually exclusive subsets
        FOR i in <1 to P>
            WHILE length(feature_subset) > length(feature_set)/P:
                REMOVE <least_significant_feature>
            UPDATE <feature_subset>
END
```

5.3.3.3 Least Absolute Shrinkage and Selection Operator (LASSO)

Least Absolute Shrinkage and Selection Operator (LASSO) (TIBSHIRANI, 1996) is a regularization method in which a penalty term is added to the likelihood function optimized in linear regression. The unconstrained OLS estimates $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ can be vulnerable to high variance, which in turn can affect inference negatively. Thus, a penalty term for the magnitude of the coefficients can control the variance. Similarly to ridge regression, in which the penalty term is the ℓ_2 norm for the $\boldsymbol{\beta}$ parameters, in LASSO the penalty is the ℓ_1 norm. The main difference is that the LASSO can yield a set of sparse solutions for the betas, making LASSO an embedded feature selection method, as the algorithm training process is done at the same time as the feature selection.

The coefficients of the LASSO regression are the solutions of the following constrained optimization problem:

$$\boldsymbol{\beta}(\boldsymbol{\lambda}) = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \frac{1}{N}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \right\}, \text{ subject to } \sum_{j=1}^{p} |\beta_j| < t \tag{5.4}$$

which is equivalent to:

$$\boldsymbol{\beta}(\boldsymbol{\lambda}) = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \frac{1}{N} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda ||\boldsymbol{\beta}||_1 \right\} \tag{5.5}$$

where λ is a free regularization parameter that controls the degree of shrinkage of the betas. Therefore, sufficiently large values for λ will effectively force some betas to be zero, producing a sparse solution for the LASSO estimator. In general, the optimal λ that minimizes the out-of-sample error is found by manually tuning through K-fold cross-validation.

For classification problems, the ℓ_1-regularization is analogously introduced to the likelihood function optimized in logistic regression, which leads to the LASSO logistic regression, whose coefficients are obtained solving the following optimization problem:

$$\hat{\boldsymbol{\beta}(\boldsymbol{\lambda})} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \frac{1}{N} \sum_{i=1}^{N} \rho_{(\boldsymbol{\beta})}(X_i, Y_i) + \lambda ||\boldsymbol{\beta}||_1 \right\} \tag{5.6}$$

where $\rho_{(\boldsymbol{\beta})}(X_i, Y_i) = -y \left\{ \sum_{j=0}^{k} \beta_j x^{(j)} + log \left[1 + exp \left(\sum_{j=0}^{k} \beta_j x^{(j)} \right) \right] \right\}$ is the likelihood function optimized to obtain the beta coefficients in equation **5.2**.

5.4 Empirical analysis

We collected daily data between January 1st, 2008 and March 1st, 2019 from firms that composed financial market indexes from seven markets, namely: United States (S&P 100 Index), United Kingdom (FTSE 100 Index), France (CAC 40 Index), Germany (DAX-30 Index), Japan (Top 50 assets from NIKKEI 225 Index), China (Top 50 assets from SSE 180 Index) and Brazil (Bovespa Index). The independent variables are the technical analysis indicators listed on tables **23** and **24** for period t, and the dependent variable is the price direction movement between periods t and $t + 1$.

Given the high number of technical indicators listed in tables **23**, which are variables already used by the scientific community, the feature selection methods described in section **5.3.3** were applied and their respective predictive performances were tested. The same procedure was reiterated considering the indicators listed in both tables **23** and **24** to verify which indicators not yet used in academic researches can potentially boost the models' explanatory power.

The collected datasets were split into four sequential and mutually exclusive subsets: the first two subsets, composed by observations between January 1st, 2008 and December 31th, 2010, were used for training and validation of the feature selection algorithms (with training-testing proportion of 75% to 25%), while the data from January 1st,

2011 to March 1st, 2019 were applied to the training and validation of the ANN models (also with training-testing proportion of 75% to 25%). For the wrapper feature selection methods, logistic regressions were fitted in the first training set with the candidate features and perform additions/removals based on its performance on the first validation set, using the accuracy as the evaluation metric.

For the LASSO the procedure was similar: the hyperparameter λ was tuned by 10-fold cross-validation in the first training set with a grid-search for the λ hyperparameter, and the value for λ that minimizes the classification error (measured by the binomial deviance) is considered to the best hyperparameter and this value of λ was used to fit the model in the validation set, and the variables for which the LASSO estimates for this model are non-zero were considered as the variable subset yielded by the feature selection.

After reaching a refined variable subset from SFFS, TS, and LASSO, the selected variables were applied in the second training dataset to fit the Deep Neural Networks; different numbers of hidden layers and dropout rates were tested to further analyze the effects of deep network architectures and degree of regularization. The optimal weights obtained in this step were finally applied in the second validation set to verify the out-of-sample predictive performance of the models, which were measured not only by accuracy but also by precision, recall, and F-Score. We tested neural networks with 3, 5, and 7 hidden layers, with the Sigmoid function as the activation function for all cases; we tested two parameter values for Dropout (0 and 0.3). For the training of the networks we used the Adam optimization algorithm (KINGMA; BA, 2014), 400 training epochs and mini-batches of size 128 for all cases. All tests were replicated for both tables **23** and **24** (Literature + Market) and only for table **23** (Literature only).

As shown in Henrique, Sobreiro and Kimura (2019), the vast majority of studies that applied machine learning methods in stock price direction prediction uses accuracy for performance measurement; however, this metric does not take into account the proportion between true positives and false positives (Type I Error) nor the proportion between true positives and false negatives (Type II Error), making the accuracy rate potentially misleading, especially when the classes are unbalanced.

Accuracy, precision, recall, and F-Score are given, respectively, by:

$$\begin{aligned} \text{Accuracy} &= \frac{TP + TN}{TP + FP + TN + FN} \\ \text{Precision} &= \frac{TP}{TP + FP} \\ \text{Recall} &= \frac{TP}{TP + FN} \\ \text{F} - \text{Score} &= \left(\frac{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}}{2} \right)^{-1} \end{aligned} \tag{5.7}$$

where TP is the number of true positives, FP is the number of false positives, TN is the number of true negatives and FN is the number of false negatives, with the positive class being associated to an increase in the stock's price between t and $t-1$ and the negative class being assigned upon a price decrease.

Precision is a metric that penalizes the Type I Error – in this case, stocks whose prices dropped when a rise was predicted; while recall yields lower values with the presence of the Type II Error – stocks that gained value but were classified as not worth buying. Therefore, the precision can be an indicator of potential interest for investors with a high degree of risk aversion, which prioritizes avoiding bad investment choices; on the other hand, the recall gives an indication of how many potentially profitable investment opportunities are being missed, which is proportionally more costly for investors with a higher appetite for risk. Finally, the F-Score – which is the harmonic mean of precision and recall – gives a conservative middle-ground between the two types of error, in the sense of yielding a high value only if both precision and recall are high, being sensible to low values from both indicators.

In a scenario where all predictions are correct, all four metrics would exhibit a perfect score of 1; however, in a realistic mixed scenario between misclassifications from both classes (*i.e.*, both uptrends predicted as price drops and downtrends predicted as price rises), the traditionally more used accuracy rate can be misleading, especially if the predictions are heavily unbalanced (concentrated in one of the two classes). In those cases, observing also the precision and the recall can reveal more details about the real quality of the predictions, as well as providing a quick overview about the model's propensity to tend itself to Type I or Type II Errors. Finally, the F-Score provides a practical way to see the average consistency of the model to both error types, which in this application would represent bad resource allocations, both when buying an overly expansive asset and when selling a holding asset for a price too small.

5.5 Results and discussion

5.5.1 Feature selection of technical analysis indicators

Concerning the "factor zoo" of technical analysis indicators described in tables **23** and **24**, the first step was to apply feature selection methods detailed in section **5.3.3** (SFFS, TS, and LASSO) for the first set of training-testing periods (using data between 2008 and 2010). In a scenario where all technical analysis indicators are equally relevant in terms of predictive power, one could expect that, on average, all columns were picked a similar number of times, such that, conversely, a non-uniform distribution on the incidence of specific indicators being more frequently chosen can be an indication of importance. In this sense, the number of time that each indicator – from both Literature researches and services used by investors to operate in the Markets – was picked by any of the three feature selection methods was aggregated across all seven analyzed financial markets, and the distribution of times chosen and its histogram are displayed, respectively, in table **25** and figure **13**.

Technical Analysis indicators	Number of times chosen
DPO	21
HULL, MFM	16
ADO, APO, BIAS, DEMA, VOLAT	15
MOM, NVI, RVI, VOLR	14
ADX, BB_BW, BWW, DMI, DSS, FORCE, MQO_BETA	13
ADL, ATRP, DIU, MASS, MQO_STD, NATR, ULTOSC	12
AVOL, BRATIO, CCI, CHOSC, CVOL, EMV, HPR, PSK, PSY, REX, RMF, ROC, STOCH_D, VAMA, VMOM	11
AR_NEG, AR_POS, CHOPPINESS, CMO, MQO_ALPHA, MQO_PRED, PVOH, RSI, VARR, VOLUME, VOOSC	10
ARATIO, CLOSE, COPP, DID, KAMA, KST, LPR, MACDH, MIDPOINT, MPP, OBV, OSCP, PVI, PVOI, STOCH_K	9
AD, ATR, CMF, DS1, NVOI, PERC_B, PVT, TP, TSI	8
AR_OSC, DISP, EMA, FR2, KC_L, KC_M, MAE_UP, PPOH, SS1, STOCH_D_SLOW, VPT	7
AB_DOWN, BB_LOW, DONCHIAN, FS1, FS2, KC_U, MAE_LOW, MFI, PC_DOWN, SAR, TRIX, WILL_R	6
BB_UP, CHAND_SHORT, DR1, FR1, MIDPRICE, OPEN, PVO, SR1, SR2, SS2, TRIMA, VWAP	5
AB_UP, CHAND_LONG, MACD, PD1, PPO, ULCER	4
MAXX, MINN, TEMA	3
PC_UP, SMA, WWS	2
WMA	1

Table 25 – Distribution of the number of times that each technical analysis indicator (Literature + Market) was chosen by the feature selection methods throughout the seven markets

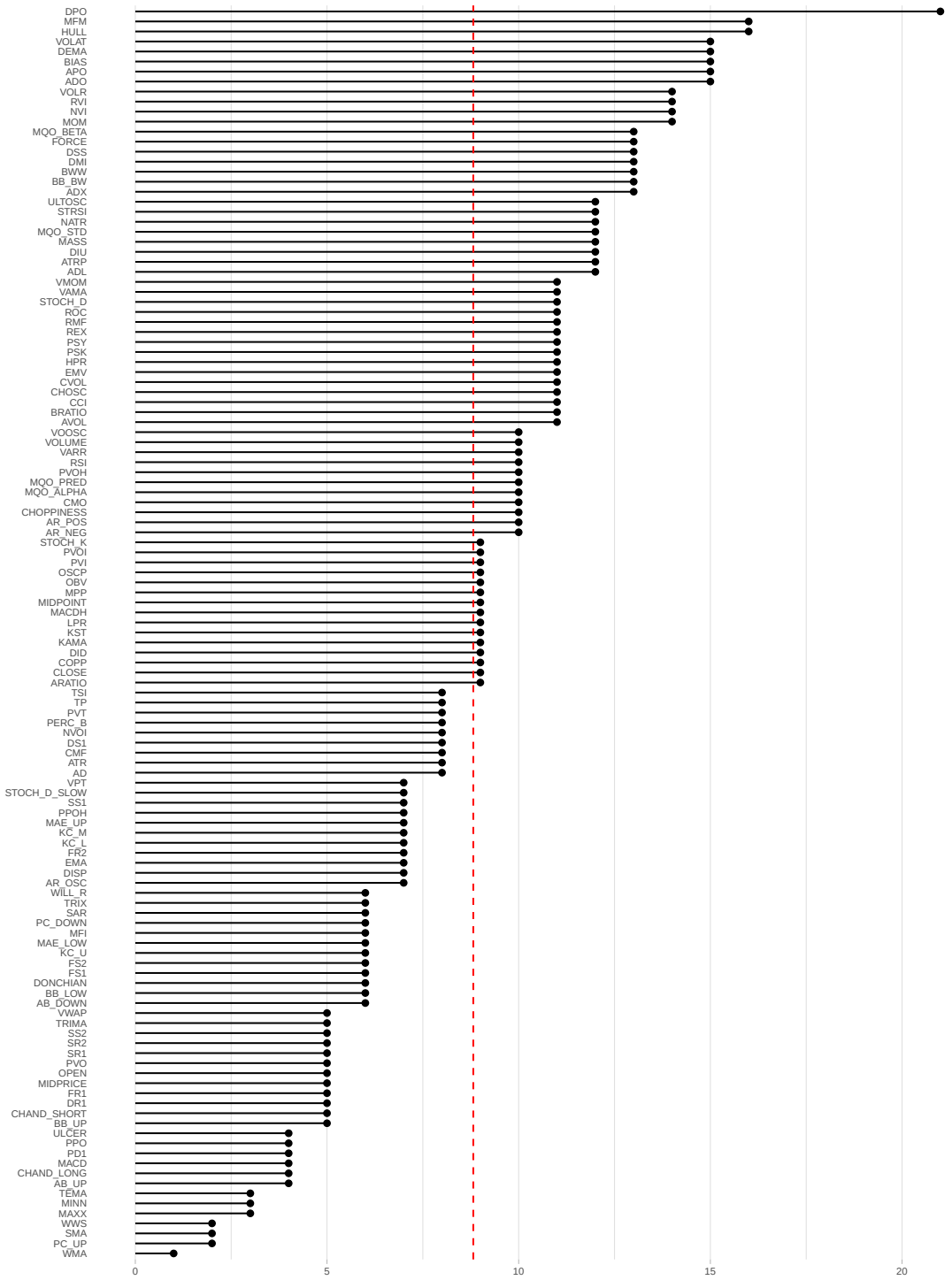

Figure 13 – Histogram of the number of times that each technical analysis indicator (Literature + Market) was chosen by the feature selection methods throughout the seven markets. The dashed vertical line indicates the average value considering all indicators.

As seen in table 25, the empirical distribution presented a fairly assymetrical behavior, with indicators like DPO (Detrended Price Oscillator), HULL (Hull Moving Average) and MFM (Money Flow Multiplier) being chosen 16 times or more, while indicators widely used in scientific researches like SMA (Simple Moving Average) and WMA

(Weighted Moving Average) were picked in a very small number of occasions. Analogously, figure **13** shows that out of the 68 technical analysis indicators that were chosen more times than the average of all 125 columns in tables **23** and **24**, 38 belong to the "market side", whilst only 30 were already being considered by academic papers. Given that many technical analysis indicators bear similar ways of calculation, formulae that combined more sources of information seemed to have been prioritized over "simpler" indicators – for instance, the Hull Moving Average is a combination of Weighted Moving Averages, and the feature selection methods, by identifying this combination as informative, probably interpreted the simpler WMA as a redundant source of information. Similar results were found using only the indicators from table **23**, as shown below in table **26** and figure **14**.

Technical Analysis indicators	Number of times chosen
ADO	16
BIAS, DIU, MACD, VOLR, VOLUME	14
CCI, DID, LPR, NVI, STOCH_K, WILL_R	13
REX, VARR	12
ARATIO, DISP, DMI, STOCH_D	11
HPR, MPP, PPO, PSY, ROC, WMA	10
MOM, MQO_BETA, OSCP, SAR, VOLAT	9
BRATIO, MFI, OBV, PVI	8
MAE_LOW	7
RSI, STOCH_D_SLOW, VAMA, VMOM	6
ATR, BB_UP, CLOSE, SMA	5
BB_LOW, KC_L, MAE_UP, MAXX, MINN	4
EMA, OPEN	3
KC_U, TRIMA	2

Table 26 – Distribution of the number of times that each technical analysis indicator (Only Literature) was chosen by the feature selection methods throughout the seven markets

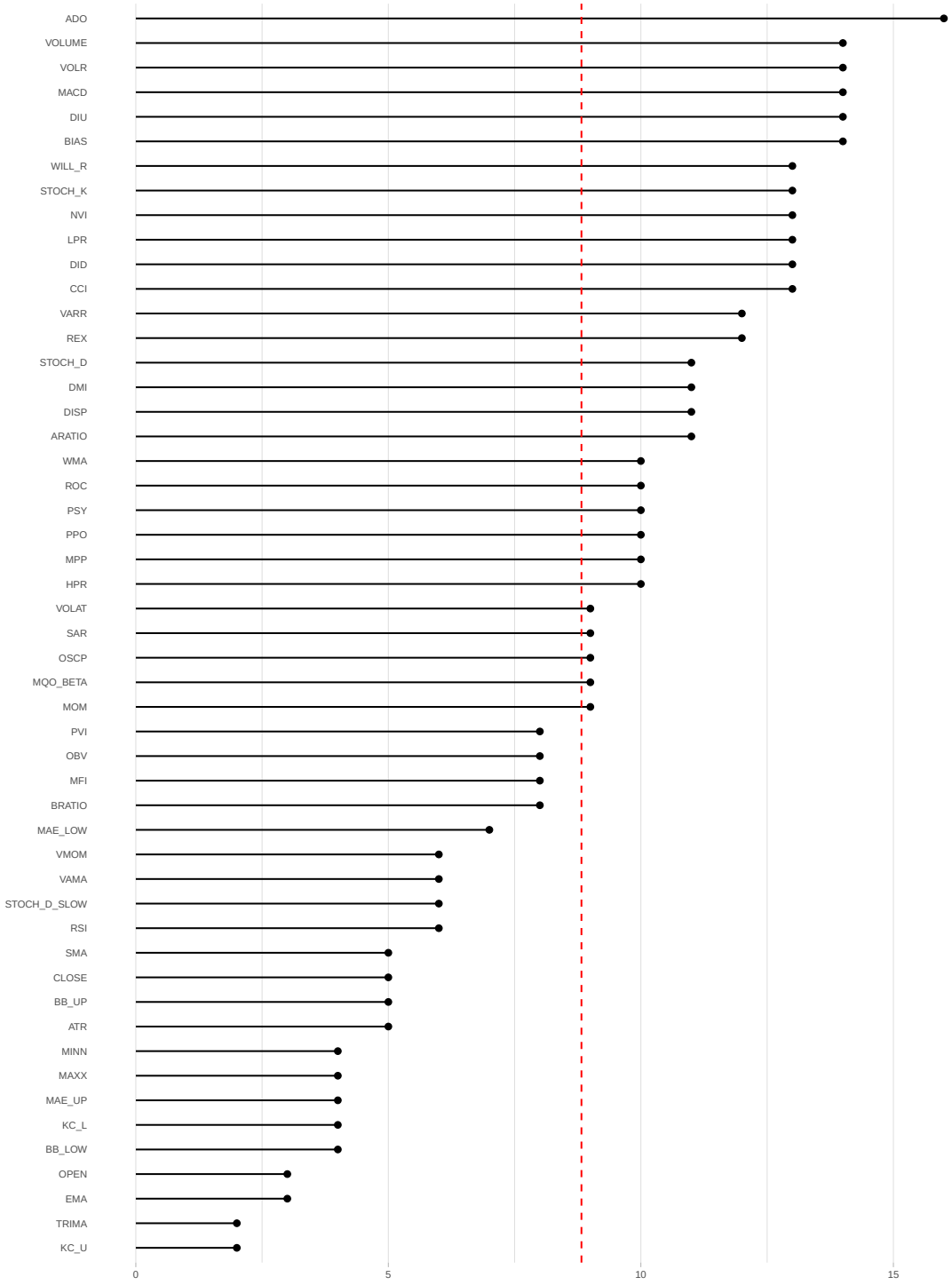

Figure 14 – Histogram of the number of times that each technical analysis indicator (Only Literature) was chosen by the feature selection methods throughout the seven markets. The dashed vertical line indicates the average value considering all indicators.

Indeed, out of the 51 technical analysis indicators used in recent scientific articles, 29 were picked more than the average, also exhibiting an assymetrical behavior in which some columns like ADO (Accumulation/Distribution Oscillator), BIAS (Bias), DIU (Directional Indicator Up) and MACD (Moving Average Convergence-Divergence) were

picked a high amount of times, while "simpler" indicators like EMA (Exponential Moving Average) and OPEN (Opening price of the day) were less picked. Just like the observed pattern for "Literature + Market", combinations of simpler indicators were identified as "informative", while the constitutes of those indicators were regarded as "redundant" – indeed, BIAS is a combination of CLOSE (Closing price) and SMA, and MACD is a difference of EMAs, all of those were chosen as separate features a small number of times in comparison.

5.5.2 Predictive performance

Technical Analysis indicators	Feature selection method	Hidden layers	Dropout	Accuracy	Precision	Recall	F-Score
Literature + market (Tables 23 and 24)	None	3	0	0.6524764	0.6573566	0.7035890	0.6796875
			0.3	0.6449702	0.6686443	0.6393739	0.6536816
		5	0	0.5076644	0.5327298	0.4924761	0.5118127
			0.3	0.6224303	0.6233794	0.7061238	0.6621767
		7	0	0.6515065	0.6549773	0.7078941	0.6804084
			0.3	0.6486179	0.6598852	0.6799308	0.6697580
	SFFS	3	0	0.5174057	0.5258766	0.8037740	0.6357850
			0.3	0.5119658	0.5262834	0.6880180	0.5963799
		5	0	0.6519704	0.6663876	0.6726080	0.6694834
			0.3	0.6496510	0.6562239	0.6961455	0.6755955
		7	0	0.5173635	0.5284836	0.7330812	0.6141918
			0.3	0.6361777	0.6360414	0.7147340	0.6730955
	TS	3	0	0.6514854	0.6757145	0.6440412	0.6594978
			0.3	0.6505366	0.6707498	0.6543414	0.6624440
		5	0	0.5150020	0.5263863	0.7432606	0.6163008
			0.3	0.5126616	0.5252839	0.7276495	0.6101243
		7	0	0.6460455	0.6692685	0.6416673	0.6551774
			0.3	0.6523078	0.6562663	0.7066468	0.6805254
	LASSO	3	0	0.6424821	0.6491427	0.6915587	0.6696797
			0.3	0.5061252	0.5232283	0.6484670	0.5791545
		5	0	0.6459612	0.6574620	0.6772753	0.6672216
			0.3	0.6487655	0.6399399	0.7540034	0.6923049
		7	0	0.5113965	0.5260547	0.6827875	0.5942605
			0.3	0.5149177	0.5264787	0.7391969	0.6149623
Literature (Table 23)	None	3	0	0.6508529	0.6647337	0.6733725	0.6690252
			0.3	0.6525819	0.6646358	0.6803331	0.6723929
		5	0	0.6163156	0.6251363	0.6690271	0.6463374
			0.3	0.5185443	0.5401176	0.5471152	0.5435938
		7	0	0.6490607	0.6648727	0.6660497	0.6654607
			0.3	0.6536572	0.6719018	0.6627102	0.6672743
	SFFS	3	0	0.5164780	0.5266530	0.7640219	0.6235101
			0.3	0.5160141	0.5274044	0.7356160	0.6143481
		5	0	0.6460877	0.6491994	0.7063249	0.6765584
			0.3	0.6457081	0.6352064	0.7609238	0.6924049
		7	0	0.5109537	0.5306907	0.5774523	0.5530849
			0.3	0.4945495	0.5267014	0.3500040	0.4205463
	TS	3	0	0.6480064	0.6601885	0.6765511	0.6682696
			0.3	0.6481751	0.6404305	0.7493764	0.6906333
		5	0	0.5127248	0.5261047	0.7070894	0.6033163
			0.3	0.5063361	0.5266191	0.5735093	0.5490649
		7	0	0.6487022	0.6655754	0.6625493	0.6640589
			0.3	0.6499251	0.6506592	0.7168665	0.6821602
	LASSO	3	0	0.6072701	0.5951537	0.7836566	0.6765196
			0.3	0.4759525	0.5240475	1.0000000	0.6877049
		5	0	0.6489763	0.6731808	0.6417076	0.6570675
			0.3	0.6514644	0.6774084	0.6394142	0.6578631
		7	0	0.5159508	0.5247721	0.8084413	0.6364284
			0.3	0.5133363	0.5280423	0.6716424	0.5912480

Table 27 – Out-of-sample prediction results for assets of S&P 100 Index

Technical Analysis indicators	Feature selection method	Hidden layers	Dropout	Accuracy	Precision	Recall	F-Score
Literature + market (Tables 23 and 24)	None	3	0	0.6162575	0.6066809	0.5946578	0.6006092
			0.3	0.6315475	0.6349206	0.5662235	0.5986076
		5	0	0.5077346	0.4928381	0.4997538	0.4962718
			0.3	0.5120946	0.4979596	0.6758986	0.5734426
		7	0	0.6258138	0.6238178	0.5764402	0.5991939
			0.3	0.6330407	0.6502732	0.5273264	0.5823817
	SFFS	3	0	0.5050469	0.4905748	0.5221566	0.5058732
			0.3	0.5068984	0.4931436	0.5843181	0.5348732
		5	0	0.6388939	0.6357815	0.5988429	0.6167596
			0.3	0.6295765	0.6551001	0.4996307	0.5668994
		7	0	0.5109001	0.4915166	0.2317824	0.3150146
			0.3	0.5147823	0.0000000	0.0000000	0.0000000
	TS	3	0	0.6364451	0.6401927	0.5725012	0.6044577
			0.3	0.6305322	0.6763100	0.4575332	0.5458150
		5	0	0.5163352	0.5012938	0.6200148	0.5543694
			0.3	0.5131697	0.4983803	0.5113245	0.5047694
		7	0	0.6195425	0.6062258	0.6160758	0.6111111
			0.3	0.6365048	0.6395126	0.5749631	0.6055224
	LASSO	3	0	0.5141253	0.4992046	0.4249138	0.4590731
			0.3	0.5147823	0.0000000	0.0000000	0.0000000
		5	0	0.6233053	0.6112131	0.6145987	0.6129012
			0.3	0.6362659	0.6570414	0.5237568	0.5828767
		7	0	0.5051663	0.4936167	0.7662482	0.6004340
			0.3	0.5096458	0.4965688	0.7660020	0.6025368
Literature (Table 23)	None	3	0	0.6371618	0.6164866	0.6674052	0.6409362
			0.3	0.6342949	0.6404605	0.5615460	0.5984128
		5	0	0.5144837	0.4998095	0.8072378	0.6173688
			0.3	0.5147823	0.0000000	0.0000000	0.0000000
		7	0	0.6275458	0.6184739	0.6065977	0.6124783
			0.3	0.6334588	0.6667226	0.4890448	0.5642264
	SFFS	3	0	0.5164546	0.5023380	0.3702610	0.4263039
			0.3	0.5071373	0.4953502	0.8392418	0.6229898
		5	0	0.6132115	0.6061582	0.5791482	0.5923455
			0.3	0.5147823	0.0000000	0.0000000	0.0000000
		7	0	0.5152004	0.5004026	0.5354505	0.5173337
			0.3	0.5147823	0.0000000	0.0000000	0.0000000
	TS	3	0	0.6158992	0.6073557	0.5894879	0.5982885
			0.3	0.5147823	0.0000000	0.0000000	0.0000000
		5	0	0.5074359	0.4946180	0.6957164	0.5781801
			0.3	0.5147823	0.0000000	0.0000000	0.0000000
		7	0	0.6382966	0.6187414	0.6632201	0.6402091
			0.3	0.6233053	0.6525097	0.4784589	0.5520915
	LASSO	3	0	0.5198590	0.5040075	0.6579271	0.5707726
			0.3	0.5147823	0.0000000	0.0000000	0.0000000
		5	0	0.6312489	0.6178675	0.6291236	0.6234447
			0.3	0.6295765	0.6605949	0.4865830	0.5603913
		7	0	0.5144239	0.4995264	0.3894633	0.4376816
			0.3	0.5147823	0.0000000	0.0000000	0.0000000

Table 28 – Out-of-sample prediction results for assets of FTSE 100 Index

Technical Analysis indicators	Feature selection method	Hidden layers	Dropout	Accuracy	Precision	Recall	F-Score
Literature + market (Tables 23 and 24)	None	3	0	0.6159981	0.6602177	0.5153374	0.5788497
			0.3	0.6350894	0.6575272	0.5998112	0.6273445
		5	0	0.5122040	0.5208290	0.5929684	0.5545625
			0.3	0.5173997	0.5220854	0.6805097	0.5908625
		7	0	0.6110440	0.6303402	0.5814063	0.6048852
			0.3	0.6381102	0.6700410	0.5778669	0.6205499
	SFFS	3	0	0.5019333	0.5217554	0.3282209	0.4029548
			0.3	0.5111165	0.5154540	0.7555451	0.6128230
		5	0	0.6448768	0.6659852	0.6149127	0.6394307
			0.3	0.6398018	0.6839333	0.5514394	0.6105813
		7	0	0.5141373	0.5198174	0.6715432	0.5860187
			0.3	0.5084582	0.5144361	0.7147239	0.5982619
	TS	3	0	0.6406477	0.6720196	0.5825861	0.6241153
			0.3	0.6349686	0.6880989	0.5252478	0.5957447
		5	0	0.5154664	0.5219653	0.6392166	0.5746712
			0.3	0.5138956	0.5171533	0.7647475	0.6170395
		7	0	0.6050024	0.6250968	0.5712600	0.5969671
			0.3	0.6214355	0.6896670	0.4740444	0.5618795
	LASSO	3	0	0.5196955	0.5276435	0.5922605	0.5580878
			0.3	0.5024166	0.5202977	0.3629070	0.4275785
		5	0	0.6220396	0.6345128	0.6177442	0.6260163
			0.3	0.6224021	0.6873107	0.4818311	0.5665141
		7	0	0.5151039	0.5182660	0.7531855	0.6140233
			0.3	0.5119623	0.5155785	0.7770175	0.6198588
Literature (Table 23)	None	3	0	0.6449976	0.6488777	0.6684757	0.6585309
			0.3	0.6367811	0.6932246	0.5214724	0.5952060
		5	0	0.5039874	0.5208008	0.3928740	0.4478816
			0.3	0.5070082	0.5187293	0.5162813	0.5175024
		7	0	0.6377477	0.6458824	0.6477112	0.6467955
			0.3	0.6428226	0.6926082	0.5438886	0.6093048
	SFFS	3	0	0.5108748	0.5186934	0.6219915	0.5656652
			0.3	0.5105123	0.5148814	0.7633318	0.6149606
		5	0	0.6071774	0.6521739	0.4990562	0.5654324
			0.3	0.4879169	0.0000000	0.0000000	0.0000000
		7	0	0.5159497	0.5169789	0.8334120	0.6381210
			0.3	0.5068874	0.5165997	0.5764512	0.5448868
	TS	3	0	0.6152731	0.6502281	0.5382256	0.5889491
			0.3	0.4879169	0.0000000	0.0000000	0.0000000
		5	0	0.5152247	0.5173579	0.7947145	0.6267213
			0.3	0.4879169	0.0000000	0.0000000	0.0000000
		7	0	0.6414936	0.6324234	0.7161397	0.6716831
			0.3	0.4879169	0.0000000	0.0000000	0.0000000
	LASSO	3	0	0.5172789	0.5204925	0.7281737	0.6070621
			0.3	0.4879169	0.0000000	0.0000000	0.0000000
		5	0	0.6384727	0.6559059	0.6184521	0.6366286
			0.3	0.4879169	0.0000000	0.0000000	0.0000000
		7	0	0.4879169	0.0000000	0.0000000	0.0000000
			0.3	0.4879169	0.0000000	0.0000000	0.0000000

Table 29 – Out-of-sample prediction results for assets of CAC 40 Index

Technical Analysis indicators	Feature selection method	Hidden layers	Dropout	Accuracy	Precision	Recall	F-Score
Literature + market (Tables 23 and 24)	None	3	0	0.5005141	0.5314402	0.5070150	0.5189403
			0.3	0.5318766	0.5351831	0.9051766	0.6726586
		5	0	0.5349614	0.5347896	0.9593614	0.6867532
			0.3	0.5326478	0.5331205	0.9695210	0.6879506
		7	0	0.4969152	0.5426357	0.3386551	0.4170390
			0.3	0.5269923	0.5334907	0.8746976	0.6627566
	SFFS	3	0	0.5254499	0.5361465	0.7929366	0.6397346
			0.3	0.5339332	0.5348327	0.9433962	0.6826536
		5	0	0.5311054	0.5576649	0.5684567	0.5630091
			0.3	0.5354756	0.5378567	0.8935656	0.6715143
		7	0	0.4871465	0.5286169	0.3217223	0.4000000
			0.3	0.4925450	0.5275311	0.4310595	0.4744409
	TS	3	0	0.5156812	0.5419532	0.5718433	0.5564972
			0.3	0.5365039	0.5369955	0.9269473	0.6800355
		5	0	0.5295630	0.5315411	0.9661345	0.6857830
			0.3	0.5336761	0.5360502	0.9100145	0.6746772
		7	0	0.5275064	0.5332752	0.8877600	0.6663036
			0.3	0.5385604	0.5393291	0.9022738	0.6751131
	LASSO	3	0	0.5329049	0.5328774	0.9801645	0.6904072
			0.3	0.5341902	0.5336323	0.9787131	0.6906794
		5	0	0.5251928	0.5369375	0.7735849	0.6338949
			0.3	0.5264781	0.5341012	0.8524432	0.6567275
		7	0	0.5336761	0.5355038	0.9230769	0.6777975
			0.3	0.5359897	0.5356172	0.9530723	0.6858139
Literature (Table 23)	None	3	0	0.5380463	0.5736900	0.5084664	0.5391126
			0.3	0.5385604	0.5393519	0.9017900	0.6749955
		5	0	0.5290488	0.5330706	0.9163038	0.6740214
			0.3	0.4922879	0.5260181	0.4499274	0.4850065
		7	0	0.5311054	0.5347639	0.9042090	0.6720604
			0.3	0.5388175	0.5388778	0.9153362	0.6783793
	SFFS	3	0	0.5218509	0.5299566	0.8858249	0.6631655
			0.3	0.5331620	0.5423272	0.7779390	0.6391097
		5	0	0.5287918	0.5603715	0.5253991	0.5423221
			0.3	0.5413882	0.5395804	0.9332366	0.6838001
		7	0	0.5326478	0.5332089	0.9671021	0.6874140
			0.3	0.5313625	0.5313625	1.0000000	0.6939735
	TS	3	0	0.5262211	0.5335731	0.8611514	0.6588932
			0.3	0.5267352	0.5352684	0.8297049	0.6507304
		5	0	0.5341902	0.5346750	0.9511369	0.6845404
			0.3	0.5313625	0.5313625	1.0000000	0.6939735
		7	0	0.5352185	0.5358626	0.9361393	0.6815780
			0.3	0.5313625	0.5313625	1.0000000	0.6939735
	LASSO	3	0	0.5329049	0.5355518	0.9109821	0.6745477
			0.3	0.5313625	0.5313625	1.0000000	0.6939735
		5	0	0.5357326	0.5359405	0.9414611	0.6830467
			0.3	0.5313625	0.5313625	1.0000000	0.6939735
		7	0	0.5262211	0.5320917	0.8984035	0.6683462
			0.3	0.5313625	0.5313625	1.0000000	0.6939735

Table 30 – Out-of-sample prediction results for assets of DAX-30 Index

Technical Analysis indicators	Feature selection method	Hidden layers	Dropout	Accuracy	Precision	Recall	F-Score
Literature + market (Tables 23 and 24)	None	3	0	0.6383089	0.6300209	0.5741144	0.6007699
			0.3	0.6301287	0.6151356	0.5869210	0.6006972
		5	0	0.4898609	0.4779222	0.8247956	0.6051781
			0.3	0.5171137	0.4922451	0.5938238	0.5382842
		7	0	0.6343049	0.6009793	0.6800182	0.6380603
			0.3	0.6288371	0.6075448	0.6128974	0.6102093
	SFFS	3	0	0.5025186	0.4781142	0.5406903	0.5074805
			0.3	0.5126362	0.4855894	0.4743869	0.4799228
		5	0	0.6374909	0.6140166	0.6334242	0.6235694
			0.3	0.6377492	0.6191700	0.6125341	0.6158342
		7	0	0.4921428	0.4778641	0.7705722	0.5899040
			0.3	0.5833297	0.5606447	0.5592189	0.5599309
	TS	3	0	0.6416240	0.6242368	0.6128974	0.6185151
			0.3	0.6365006	0.6348923	0.5486830	0.5886480
		5	0	0.4956301	0.4783010	0.7057221	0.5701706
			0.3	0.5140569	0.4873412	0.4842870	0.4858093
		7	0	0.6273303	0.6047526	0.6171662	0.6108963
			0.3	0.6334008	0.6052476	0.6515895	0.6275642
	LASSO	3	0	0.5044130	0.4778847	0.4916440	0.4846667
			0.3	0.4877083	0.4767217	0.8267938	0.6047500
		5	0	0.6288802	0.6062222	0.6194369	0.6127583
			0.3	0.6358548	0.6103234	0.6411444	0.6253544
		7	0	0.5045421	0.4791841	0.5206176	0.4990423
			0.3	0.5102252	0.4799913	0.3987284	0.4356023
Literature (Table 23)	None	3	0	0.6332716	0.6050059	0.6520436	0.6276447
			0.3	0.6413226	0.6268586	0.6011807	0.6137512
		5	0	0.4760839	0.4741583	0.9657584	0.6360401
			0.3	0.5259827	0.0000000	0.0000000	0.0000000
		7	0	0.6340466	0.6029026	0.6678474	0.6337154
			0.3	0.6410643	0.6306579	0.5859219	0.6074674
	SFFS	3	0	0.5091058	0.4835488	0.5232516	0.5026173
			0.3	0.4930469	0.4764861	0.7039964	0.5683176
		5	0	0.6247901	0.6213387	0.5336966	0.5741926
			0.3	0.6333577	0.6000803	0.6791099	0.6371538
		7	0	0.5118181	0.4866010	0.5425976	0.5130760
			0.3	0.5259827	0.0000000	0.0000000	0.0000000
	TS	3	0	0.6223361	0.6015611	0.6019982	0.6017796
			0.3	0.5259827	0.0000000	0.0000000	0.0000000
		5	0	0.5051449	0.4832919	0.6357856	0.5491488
			0.3	0.5259827	0.0000000	0.0000000	0.0000000
		7	0	0.6392130	0.6233815	0.6034514	0.6132546
			0.3	0.6313342	0.5936042	0.7047230	0.6444085
	LASSO	3	0	0.5017006	0.4783709	0.5664850	0.5187126
			0.3	0.5259827	0.0000000	0.0000000	0.0000000
		5	0	0.6370603	0.6432697	0.5260672	0.5787948
			0.3	0.6330133	0.5936699	0.7155313	0.6489292
		7	0	0.5139278	0.4845815	0.3996367	0.4380289
			0.3	0.5259827	0.0000000	0.0000000	0.0000000

Table 31 – Out-of-sample prediction results for top 50 assets of NIKKEI 225 Index

Technical Analysis indicators	Feature selection method	Hidden layers	Dropout	Accuracy	Precision	Recall	F-Score
Literature + market (Tables 23 and 24)	None	3	0	0.6277843	0.6695116	0.4116018	0.5097933
			0.3	0.6315537	0.6120755	0.5910217	0.6013644
		5	0	0.5275170	0.4954357	0.2610407	0.3419244
			0.3	0.5153862	0.4845361	0.4795219	0.4820160
		7	0	0.6331300	0.6710526	0.4311325	0.5249800
			0.3	0.6275786	0.6094493	0.5790701	0.5938714
	SFFS	3	0	0.5287506	0.4975482	0.2218336	0.3068548
			0.3	0.5290247	0.4988098	0.3359569	0.4014980
		5	0	0.6434789	0.6582713	0.5028422	0.5701537
			0.3	0.6412172	0.6328866	0.5643492	0.5966561
		7	0	0.5301898	0.5007448	0.2939805	0.3704656
			0.3	0.5377973	0.5118541	0.3681679	0.4282808
	TS	3	0	0.6384072	0.6467864	0.5089637	0.5696574
			0.3	0.6331985	0.6179828	0.5760093	0.5962583
		5	0	0.5348503	0.5161855	0.1719866	0.2580081
			0.3	0.5307381	0.5025473	0.2012826	0.2874389
		7	0	0.6164759	0.5874223	0.6194432	0.6030079
			0.3	0.6303201	0.6701461	0.4210756	0.5171858
	LASSO	3	0	0.5192927	0.4844354	0.3470340	0.4043818
			0.3	0.5298472	0.5001391	0.2620609	0.3439174
		5	0	0.6247687	0.5889145	0.6689987	0.6264074
			0.3	0.6349119	0.6715116	0.4376913	0.5299568
		7	0	0.5264889	0.4922027	0.2208133	0.3048596
			0.3	0.5348503	0.5131020	0.2111937	0.2992256
Literature (Table 23)	None	3	0	0.6377904	0.6589992	0.4760239	0.5527630
			0.3	0.6438215	0.6523252	0.5193121	0.5782683
		5	0	0.5336851	0.5127860	0.1665938	0.2514851
			0.3	0.5332054	0.5068644	0.2690570	0.3515186
		7	0	0.6345693	0.6267617	0.5509401	0.5864102
			0.3	0.6294976	0.6629339	0.4314240	0.5226912
	SFFS	3	0	0.5319032	0.5053100	0.2149832	0.3016360
			0.3	0.5334795	0.5109489	0.1836467	0.2701833
		5	0	0.6227126	0.6177083	0.5185833	0.5638222
			0.3	0.6318278	0.6787515	0.4120391	0.5127880
		7	0	0.5205264	0.4870017	0.3686052	0.4196117
			0.3	0.5336166	0.5069307	0.2984988	0.3757453
	TS	3	0	0.6239463	0.6152685	0.5344702	0.5720303
			0.3	0.6445754	0.6550065	0.5158140	0.5771363
		5	0	0.5254609	0.4917733	0.2744498	0.3522919
			0.3	0.5339593	0.5072602	0.3105961	0.3852830
		7	0	0.6397094	0.6143427	0.6280426	0.6211171
			0.3	0.6421081	0.6560655	0.5021134	0.5688573
	LASSO	3	0	0.5347817	0.5115763	0.2350969	0.3221490
			0.3	0.5297786	0.0000000	0.0000000	0.0000000
		5	0	0.6126379	0.5707100	0.7111208	0.6332252
			0.3	0.6361456	0.6584967	0.4699023	0.5484392
		7	0	0.5303954	0.5014254	0.2307244	0.3160311
			0.3	0.5297786	0.0000000	0.0000000	0.0000000

Table 32 – Out-of-sample prediction results for top 50 assets of SSE 180 Index

Technical Analysis indicators	Feature selection method	Hidden layers	Dropout	Accuracy	Precision	Recall	F-Score
Literature + market (Tables 23 and 24)	None	3	0	0.6286628	0.6151003	0.6408627	0.6277173
			0.3	0.6420676	0.6247080	0.6694589	0.6463097
		5	0	0.5035511	0.4917489	0.4848835	0.4882920
			0.3	0.5150275	0.5063420	0.2882727	0.3673845
		7	0	0.6327548	0.6170329	0.6543424	0.6351402
			0.3	0.6391515	0.6126515	0.7105719	0.6579886
	SFFS	3	0	0.5097597	0.4973416	0.3332370	0.3990775
			0.3	0.5124877	0.5013039	0.3886963	0.4378762
		5	0	0.6335074	0.6208535	0.6415367	0.6310257
			0.3	0.6415032	0.6264641	0.6591566	0.6423947
		7	0	0.5146042	0.5151376	0.1081263	0.1787363
			0.3	0.5155919	0.5102329	0.2088388	0.2963722
	TS	3	0	0.6399981	0.6254592	0.6556904	0.6402181
			0.3	0.6443723	0.6252105	0.6790872	0.6510361
		5	0	0.5128169	0.5024146	0.2804737	0.3599852
			0.3	0.5114999	0.0000000	0.0000000	0.0000000
		7	0	0.6303090	0.6140303	0.6548238	0.6337713
			0.3	0.6421147	0.6198533	0.6914115	0.6536799
	LASSO	3	0	0.5099948	0.4977578	0.3419988	0.4054332
			0.3	0.5114999	0.0000000	0.0000000	0.0000000
		5	0	0.6311086	0.6040763	0.7105719	0.6530107
			0.3	0.6429143	0.6236064	0.6786058	0.6499447
		7	0	0.5103711	0.4980347	0.2927980	0.3687849
			0.3	0.5109826	0.4993266	0.3926439	0.4396055
Literature (Table 23)	None	3	0	0.6313908	0.6147475	0.6574235	0.6353697
			0.3	0.6387752	0.6358707	0.6096669	0.6224931
		5	0	0.5090071	0.4966587	0.3792605	0.4300923
			0.3	0.5114999	0.0000000	0.0000000	0.0000000
		7	0	0.6426791	0.6352439	0.6306567	0.6329420
			0.3	0.6434787	0.6242253	0.6787984	0.6503690
	SFFS	3	0	0.5105592	0.4967469	0.1470248	0.2268945
			0.3	0.5114999	0.0000000	0.0000000	0.0000000
		5	0	0.6264992	0.6107036	0.6493356	0.6294274
			0.3	0.6424439	0.6212121	0.6868862	0.6524005
		7	0	0.5091012	0.4973368	0.4585018	0.4771304
			0.3	0.5114999	0.0000000	0.0000000	0.0000000
	TS	3	0	0.6331781	0.6153160	0.6645484	0.6389853
			0.3	0.6336485	0.5966218	0.7720008	0.6730745
		5	0	0.5023752	0.4911140	0.5161756	0.5033330
			0.3	0.5114999	0.0000000	0.0000000	0.0000000
		7	0	0.6316731	0.6067697	0.6990179	0.6496354
			0.3	0.6359061	0.6004863	0.7609282	0.6712532
	LASSO	3	0	0.5154979	0.5083975	0.2477373	0.3331391
			0.3	0.5114999	0.0000000	0.0000000	0.0000000
		5	0	0.6421617	0.6210774	0.6860196	0.6519352
			0.3	0.6430083	0.6211438	0.6901598	0.6538356
		7	0	0.5105122	0.4985938	0.3584633	0.4170728
			0.3	0.5114999	0.0000000	0.0000000	0.0000000

Table 33 – Out-of-sample prediction results for assets of Bovespa Index

As shown in tables 27 to 33, the predictive performance for all seven markets and all 48 combinations of hyperparameters in each market (Literature + Market/Only literature, technical analysis indicators chosen by feature selection, number of hidden layers and dropout rate), were basically around two key values: the accuracy of all cases

were concentrated around 50% and 65%. The first value is consistent with the scenario postulated by the Efficient Markets Hypothesis in its weak form, which implies that no strategy can sistematically beat the Random Walk simply using past data; indeed, a fairly large proportion of the results estimated in this study converged to this state, regardless of the deepness of the neural networks, the filtering of potentially more informative technical analysis indicators and different settings of regularization. In this sense, the emergence of accuracy rates close to 50% is a sign that financial markets do tend indeed towards the equilibrium in which on average there's no significant abnormal gains over the market. In special, all results for the German market (table **30**) gravitated around 50% of accuracy

However, on the other hand, parallel to the theoretically intuitive accuracy of 50%, many cases converged to a kind of "strange attractor" of 65% of accuracy, which is, in turn, a measure that argues favorably towards the existence of profit margins above the market level. For all cases in which the accuracy rate did not lie at the surroundings of 50%, they converges systematically to 65%; computational experiments made using more training epochs showed that those cases do indeed reach a species of "stationary state" in 65%. This pattern appeared in all analyzed markets (except for the German one) and for both sources of information (technical analysis indicators from Literature + Market or only from Literature). Those two scenarios were observed for all feature selection methods (including the "None" case, in which all columns were used for the training procedures).

Moreover, the emergence of the 65% accuracy value did not seem to have a clearly distinguishable pattern amongst its occurrences: regarding the number of hidden layers in the networks, it can be seen that a 65% accuracy tend to jointly appear for the cases with 3 and 7 hidden layers, or only for the case with 5 hidden layers, with few exceptions for the case with 3 hidden layers but without dropout. This implies that the empirical behavior of cases with 3 and 7 hidden layers bear similarities, but potentially large differences in comparison to the case with 5 hidden layers. This finding can be further analyzed in other financial applications of deep neural networks to better understand the potentially chaotic behavior of the number of hidden layers in this knowledge field.

The effect of the dropout regularization in the neural networks also was pretty heterogeneous: in comparison to the case without dropout, turning off 30% of the neurons at each training epoch seemed to make a small effect (sometimes positive and sometimes negative) in the out-of-sample accuracy rates, while in some cases the performance metrics had a notable worsening with the presence of dropout, especially for the case with 3 hidden layers. In this sense, the effectiveness of dropout as a regularization tool appeared more evident for deeper networks, being potentially useful to control the complexity of the models when the number of hidden layers grow; on the other hand, in shallower network structures this mechanism can actually hinder the classification quality and lead to sub-optimal financial decision making.

Concerning the other classification metrics, on average they were close to the accuracy rate, as the predictions were approximately balanced for the majority of hyper-parameter combinations. Throughout the 336 combinations across the seven markets, in 40 the predictions yielded only one class – 33 cases that predicted that the prices would only drop, and 7 cases predicting that the prices would only rise. In a "only drop" case, the precision and recall would be zero, as no predictions were made for the "positive class"; similarly, in a "only rise" case, the recall would be equal to one, as a false negative would not exist since no predictions were made towards the "negative class".

5.5.3 Profitability of strategies and transaction costs

Besides the classification metrics discussed in the previous sections, we evaluated the profitability of the strategies based on the predictions made by the deep learning models and the maximum value for the transaction cost in the respective market for the machine learning algorithms to be able to break-even (TC_0) and to beat the Buy-and-Hold strategy (TC_{BH}). The profitability of the Buy-and-Hold was computed as the average profitability of buying all assets of the respective market at the first day of the out-of-sample testing period and selling them at the last day of this period – which is equivalent to the gains of the uniform ($\frac{1}{N}$) portfolio during this period. The results are displayed in tables **34** to **40**.

Technical Analysis indicators	Feature selection method	Hidden layers	Dropout	Strategy profitability	Number of transactions	TC_0	TC_{BH}
Literature + market (Tables 23 and 24)	None	3	0	76.3370	91	0.7891382	0.4918951
			0.3	84.7517	96	0.8742593	0.5695643
		5	0	-26.2018	125	-0.4878055	-0.9810900
			0.3	0.2977	213	-0.1306670	-0.3258851
		7	0	73.7097	89	0.8543226	0.5109819
			0.3	84.1676	94	0.8464601	0.5594070
	SFFS	3	0	-33.8689	60	-0.7833762	-1.1259025
			0.3	-93.9103	127	-0.7350088	-0.9510824
		5	0	71.9673	98	0.7003559	0.4224501
			0.3	73.8737	87	0.7715379	0.4649007
		7	0	-91.4242	72	-1.3960984	-1.6452585
			0.3	17.7903	210	-0.0991440	-0.3066543
	TS	3	0	78.5373	96	0.7697605	0.4994064
			0.3	80.4771	89	0.9240393	0.6027553
		5	0	-86.9147	73	-1.3871419	-1.7970031
			0.3	-42.7049	105	-0.4539531	-0.6951643
		7	0	83.7943	99	0.8832354	0.5832598
			0.3	77.6707	94	0.8058476	0.5031705
	LASSO	3	0	61.9372	219	0.2759900	0.1193453
			0.3	-53.5714	91	-0.6422659	-1.0093393
		5	0	79.2360	94	0.8422404	0.5369936
			0.3	65.0487	85	0.8121663	0.4136248
		7	0	-43.1031	132	-0.3601209	-0.5558035
			0.3	-31.4264	66	-0.6612784	-1.0125791
Literature (Table 23)	None	3	0	76.9507	92	0.7869704	0.5008944
			0.3	72.1548	93	0.7377142	0.4497348
		5	0	-6.3772	223	-0.0623490	-0.2078810
			0.3	-18.8899	123	-0.1465315	-0.4879969
		7	0	77.3609	90	0.8154493	0.5065409
			0.3	75.7934	93	0.7555012	0.4781839
	SFFS	3	0	-86.7935	90	-1.5201857	-2.0062484
			0.3	-87.8024	71	-1.3302887	-1.6687141
		5	0	76.9180	94	0.8084604	0.5005134
			0.3	60.8778	91	0.6999339	0.3442414
		7	0	-35.1803	155	-0.5275478	-1.1137675
			0.3	-1.5872	76	0.0849567	-0.6821820
	TS	3	0	71.1070	89	0.7297645	0.4288793
			0.3	59.3654	86	0.7342274	0.3477410
		5	0	-43.1454	111	-0.4110883	-0.6155696
			0.3	-26.8233	70	-0.4755861	-0.8923210
		7	0	85.1358	91	0.8732660	0.5883143
			0.3	58.7055	93	0.6590593	0.3095681
	LASSO	3	0	-26.4434	195	-0.1963877	-0.3705473
			0.3	-130.7093	1	-130.7092565	-160.3766018
		5	0	79.8481	89	0.8421356	0.5438817
			0.3	77.9161	93	0.7701583	0.4981121
		7	0	-111.1510	89	-1.4438992	-1.7723011
			0.3	-20.1832	71	-0.1628726	-0.9002705
Buy-and-Hold strategy profitability over the out-of-sample period: 31.04701							

Table 34 – Trading profitability and transaction costs of machine learning algorithms for assets of S&P 100 Index

Technical Analysis indicators	Feature selection method	Hidden layers	Dropout	Strategy profitability	Number of transactions	TC_0	TC_{BH}
Literature + market (Tables 23 and 24)	None	3	0	-1231.6579	204	-6.6718986	-2.5560575
			0.3	-876.1521	197	-4.9394700	-0.7267911
		5	0	-679.8966	222	-3.4837835	0.2314074
			0.3	-342.2837	180	-2.1448187	2.4721083
		7	0	-818.2142	198	-4.2889865	-0.4163683
			0.3	-1022.4958	191	-6.1411641	-1.6068589
	SFFS	3	0	-728.0858	221	-3.7490086	0.0032730
			0.3	-483.8384	200	-2.6855331	1.3802731
		5	0	-1001.1223	197	-5.6457115	-1.4699595
			0.3	-977.1600	186	-5.9200385	-1.4305391
		7	0	-872.9466	148	-5.6816792	-0.8380714
			0.3	0.0000	0	0.0000000	0.0000000
	TS	3	0	-905.6980	197	-4.8306573	-0.8758480
			0.3	-953.5090	182	-5.9360502	-1.2925984
		5	0	-795.1669	213	-4.1427479	-0.2757212
			0.3	-294.7987	203	-1.7610705	2.4298399
		7	0	-1010.0929	206	-5.0971707	-1.3522417
			0.3	-1093.3283	198	-6.0597130	-1.9353257
	LASSO	3	0	-720.3165	209	-4.0450618	-0.0097593
			0.3	0.0000	0	0.0000000	0.0000000
		5	0	-558.4443	202	-2.8833892	0.9546062
			0.3	-1102.4661	190	-6.3286272	-2.0084098
		7	0	-1068.2750	158	-8.1302902	-2.5532919
			0.3	-486.8248	152	-4.0796212	1.9143410
Literature (Table 23)	None	3	0	-880.1103	200	-4.5685387	-0.7427110
			0.3	-958.1202	194	-5.4884712	-1.2621897
		5	0	-1089.2449	135	-8.6373622	-2.9405701
			0.3	0.0000	0	0.0000000	0.0000000
		7	0	-951.0772	205	-5.1854940	-0.9388410
			0.3	-1067.3949	187	-6.1045717	-1.7920286
	SFFS	3	0	-772.5879	203	-3.8355957	-0.1945503
			0.3	-1114.6619	122	-10.7339345	-3.6361880
		5	0	-630.8446	205	-3.5275533	0.6870135
			0.3	0.0000	0	0.0000000	0.0000000
		7	0	-261.1525	209	-1.4958135	2.3081992
			0.3	0.0000	0	0.0000000	0.0000000
	TS	3	0	-759.3180	205	-3.6857697	-0.0713111
			0.3	0.0000	0	0.0000000	0.0000000
		5	0	-376.9710	182	-2.2109191	2.3533864
			0.3	0.0000	0	0.0000000	0.0000000
		7	0	-1007.0923	198	-5.4493672	-1.4586016
			0.3	-1064.7026	182	-6.7676923	-1.9861086
	LASSO	3	0	-846.9454	183	-4.9279140	-0.8298481
			0.3	0.0000	0	0.0000000	0.0000000
		5	0	-1138.4845	203	-6.0562728	-2.0636931
			0.3	-1013.1588	186	-5.8574347	-1.4905412
		7	0	-274.0323	206	-1.4215544	2.5372880
			0.3	0.0000	0	0.0000000	0.0000000
Buy-and-Hold strategy profitability over the out-of-sample period: −34.75736							

Table 35 – Trading profitability and transaction costs of machine learning algorithms for assets of FTSE 100 Index

Technical Analysis indicators	Feature selection method	Hidden layers	Dropout	Strategy profitability	Number of transactions	TC_0	TC_{BH}
Literature + market (Tables 23 and 24)	None	3	0	30.9116	88	0.3521103	0.2785103
			0.3	19.7555	80	0.2499292	0.1836907
		5	0	-17.9450	128	-0.1484100	-0.2002564
			0.3	-11.6289	73	-0.2260126	-0.3462042
		7	0	30.9218	101	0.2948034	0.2257045
			0.3	31.4456	83	0.4129977	0.3582687
	SFFS	3	0	-6.6505	93	-0.1185191	-0.2590703
			0.3	-9.4849	55	-1.0847313	-0.9257653
		5	0	22.9734	82	0.2824456	0.2045268
			0.3	28.5655	90	0.3412015	0.2763016
		7	0	-21.7477	109	-0.2165141	-0.2839465
			0.3	5.6222	72	0.0772116	-0.0297049
	TS	3	0	28.8666	94	0.3341805	0.2667978
			0.3	26.2245	82	0.4768511	0.4215231
		5	0	-11.0071	99	-0.1324484	-0.2043627
			0.3	0.3697	42	-0.4944846	-0.5041867
		7	0	32.5566	96	0.3535969	0.2933230
			0.3	31.1308	77	0.3667134	0.3063833
	LASSO	3	0	-0.8673	120	0.0043756	-0.0613975
			0.3	-0.8773	72	-0.0132984	-0.1185589
		5	0	24.0689	100	0.2615305	0.1867346
			0.3	37.7236	78	0.4752711	0.4129187
		7	0	-7.6266	63	-0.2898908	-0.3378850
			0.3	-9.6566	50	-1.1598411	-1.0225893
Literature (Table 23)	None	3	0	23.7956	84	0.2935277	0.2220890
			0.3	19.3818	87	0.2240489	0.1636593
		5	0	2.7511	99	0.0322629	-0.0640133
			0.3	5.6516	72	0.0830590	-0.0321532
		7	0	29.1989	92	0.3373511	0.2611744
			0.3	31.7185	88	0.5304779	0.4733908
	SFFS	3	0	-8.7979	97	-0.0866904	-0.1614548
			0.3	-0.8606	40	-0.7441883	-0.7401627
		5	0	23.4107	92	0.2448115	0.1647427
			0.3	0.0000	0	0.0000000	0.0000000
		7	0	-23.2653	75	-0.4770168	-0.5796258
			0.3	-9.5067	86	-0.1387403	-0.2203134
	TS	3	0	28.9731	106	0.2853326	0.2142251
			0.3	0.0000	0	0.0000000	0.0000000
		5	0	-9.2055	58	-0.5575514	-0.5791151
			0.3	0.0000	0	0.0000000	0.0000000
		7	0	20.1151	80	0.2549876	0.1755498
			0.3	0.0000	0	0.0000000	0.0000000
	LASSO	3	0	-4.3156	74	-0.0990696	-0.1910703
			0.3	0.0000	0	0.0000000	0.0000000
		5	0	27.1119	87	0.3367944	0.2611685
			0.3	0.0000	0	0.0000000	0.0000000
		7	0	0.0000	0	0.0000000	0.0000000
			0.3	0.0000	0	0.0000000	0.0000000
Buy-and-Hold strategy profitability over the out-of-sample period: 7.426471							

Table 36 – Trading profitability and transaction costs of machine learning algorithms for assets of CAC 40 Index

Technical Analysis indicators	Feature selection method	Hidden layers	Dropout	Strategy profitability	Number of transactions	TC_0	TC_{BH}
Literature + market (Tables 23 and 24)	None	3	0	-7.0142	108	-0.0999890	0.3064287
			0.3	-95.3067	30	-51.1154396	-51.0662157
		5	0	-111.6834	19	-55.6554314	-56.4373914
			0.3	-118.0445	15	-50.0377091	-49.1479528
		7	0	-57.8716	116	-0.5213857	-0.1559433
			0.3	-58.3453	28	-50.3453524	-50.3112660
	SFFS	3	0	-116.7232	26	-53.7583144	-53.6025854
			0.3	-116.6538	14	-57.9246314	-59.4331524
		5	0	5.3684	107	0.0982206	0.5336711
			0.3	-74.4657	13	-25.3718634	-31.0785939
		7	0	-50.3786	119	-0.4876987	-0.1002714
			0.3	-57.6454	88	-0.6801124	-0.2187701
	TS	3	0	-32.4863	114	-0.2170124	0.3989707
			0.3	-47.2818	28	-56.1255015	-58.3279360
		5	0	-101.1019	9	-75.3873487	-80.9304147
			0.3	-128.2447	15	-79.7961146	-91.0733620
		7	0	-37.3058	31	-48.4516385	-48.0509990
			0.3	-69.3260	28	-29.4667014	-35.9161621
	LASSO	3	0	-118.7176	13	-31.7819511	-41.8065806
			0.3	-117.4401	13	-50.3404434	-49.7330953
		5	0	-24.7099	81	-0.0299535	0.4779868
			0.3	-57.4242	22	-51.2784063	-51.4561059
		7	0	-109.0390	15	-79.5897300	-90.8774624
			0.3	-116.5046	11	-59.4268786	-62.0856623
Literature (Table 23)	None	3	0	21.9347	101	0.0947670	0.4276121
			0.3	-52.9348	12	-29.9100951	-37.5991318
		5	0	-119.4156	42	-2.6834940	-2.8865724
			0.3	-49.0859	91	-0.5618605	-0.1290446
		7	0	-66.6646	29	-57.8594311	-60.1799083
			0.3	-81.8274	19	-38.8543589	-45.0517977
	SFFS	3	0	-124.6436	39	-45.9953938	-43.6830513
			0.3	-45.0745	71	-40.0923523	-36.5127900
		5	0	-12.4590	115	-0.1727223	0.1668287
			0.3	-80.3769	16	-53.0410738	-55.4448798
		7	0	-126.0814	18	-37.6227743	-33.7698585
			0.3	-129.2861	1	-129.2861294	-91.7557131
	TS	3	0	-75.2838	26	-26.9383323	-32.0525803
			0.3	-71.8876	29	-50.4763194	-50.5510155
		5	0	-103.2904	17	-70.8163214	-80.5473919
			0.3	-129.2861	1	-129.2861294	-91.7557131
		7	0	-65.2524	17	-40.2059036	-47.0944709
			0.3	-129.2861	1	-129.2861294	-91.7557131
	LASSO	3	0	-108.2516	35	-4.9690958	-5.7337564
			0.3	-129.2861	1	-129.2861294	-91.7557131
		5	0	-119.3640	27	-36.5592164	-41.9525921
			0.3	-129.2861	1	-129.2861294	-91.7557131
		7	0	-91.3948	43	-43.9058699	-41.1566724
			0.3	-129.2861	1	-129.2861294	-91.7557131
Buy-and-Hold strategy profitability over the out-of-sample period: 13.75618							

Table 37 – Trading profitability and transaction costs of machine learning algorithms for assets of DAX-30 Index

Technical Analysis indicators	Feature selection method	Hidden layers	Dropout	Strategy profitability	Number of transactions	TC_0	TC_{BH}
Literature + market (Tables 23 and 24)	None	3	0	7803.1598	89	77.4127800	64.7076600
			0.3	7920.0890	93	88.3608673	74.6872453
		5	0	-5529.4543	51	-738.2236987	-904.1374660
			0.3	-2406.1468	151	-19.3880018	-28.6744652
		7	0	6946.7547	90	70.3862924	56.6876955
			0.3	7288.4576	97	74.6892982	61.4661992
	SFFS	3	0	-4673.4202	92	-82.4101343	-106.5526008
			0.3	-2840.2886	138	-23.0337944	-33.5978173
		5	0	7205.2919	93	69.5090016	56.9625919
			0.3	7393.7619	87	84.1624914	69.9723735
		7	0	-6433.7873	43	-5087.6461443	-6156.3638849
			0.3	1778.6500	191	9.6339904	1.7947592
	TS	3	0	7240.3170	91	70.0058116	57.5585390
			0.3	8157.9939	78	94.5646009	80.0575743
		5	0	-4951.7845	79	-629.9660252	-832.6242100
			0.3	-2842.2710	137	-21.3855507	-30.9577642
		7	0	7431.5764	92	79.4053043	65.0827782
			0.3	6658.2396	90	68.4016866	54.5935461
	LASSO	3	0	-2633.1845	128	-24.4690302	-36.4269438
			0.3	-5750.1541	53	-320.4699880	-423.5817771
		5	0	7685.6084	98	80.9512954	67.6100672
			0.3	7178.1310	92	71.1044992	58.2038033
		7	0	-3598.0590	134	-35.1086230	-48.4995211
			0.3	-4307.3004	84	-49.9346334	-63.9898179
Literature (Table 23)	None	3	0	6986.1535	81	91.9816215	74.8566646
			0.3	7116.7184	92	69.7707914	57.0395886
		5	0	-7329.1988	9	-4835.3665192	-5758.0627798
			0.3	0.0000	0	0.0000000	0.0000000
		7	0	7727.0345	87	83.7075398	69.7626794
			0.3	7345.8764	84	72.4738603	59.9154228
	SFFS	3	0	-2425.7614	128	-18.3464355	-29.3510175
			0.3	-3759.1445	66	-136.4005100	-189.0367370
		5	0	7066.3257	92	72.9686253	59.8742364
			0.3	6722.3498	93	66.8265308	53.3423158
		7	0	-3214.7343	142	-24.9785228	-35.0148337
			0.3	0.0000	0	0.0000000	0.0000000
	TS	3	0	6952.5557	96	69.0497759	56.3109355
			0.3	0.0000	0	0.0000000	0.0000000
		5	0	-4545.8384	132	-38.3434449	-48.6162344
			0.3	0.0000	0	0.0000000	0.0000000
		7	0	7665.0296	83	84.9334961	71.1548347
			0.3	6415.0398	89	70.5765688	56.1770296
	LASSO	3	0	-6435.8059	89	-107.3741112	-127.7128895
			0.3	0.0000	0	0.0000000	0.0000000
		5	0	8380.3421	83	94.5106991	80.9789519
			0.3	5719.4074	85	60.6517800	46.9317639
		7	0	633.1867	122	7.3991011	-5.8053118
			0.3	0.0000	0	0.0000000	0.0000000
Buy-and-Hold strategy profitability over the out-of-sample period: 1396.319							

Table 38 – Trading profitability and transaction costs of machine learning algorithms for the top 50 assets of NIKKEI 225 Index

Technical Analysis indicators	Feature selection method	Hidden layers	Dropout	Strategy profitability	Number of transactions	TC_0	TC_{BH}
Literature + market (Tables 23 and 24)	None	3	0	19.2356	82	0.1900204	0.0866386
			0.3	23.8928	106	0.2329831	0.1130022
		5	0	-21.5139	65	-0.7373031	-0.7669122
			0.3	-22.6591	131	-0.5747013	-0.5208803
		7	0	22.2560	84	0.2135017	0.0969307
			0.3	24.8313	106	0.2217887	0.1097081
	SFFS	3	0	-0.9213	58	-0.1404246	-0.2672765
			0.3	9.0263	100	-0.1058450	-0.1382983
		5	0	23.2658	101	0.1868674	0.0908159
			0.3	22.9580	91	0.2302031	0.1035160
		7	0	4.1075	85	0.0357258	-0.0897997
			0.3	7.7811	117	0.0648242	-0.0461792
	TS	3	0	22.9278	92	0.2105049	0.1522983
			0.3	24.3774	99	0.2448870	0.2033140
		5	0	2.1913	57	0.0881151	-0.2479354
			0.3	-1.1676	87	0.3269734	-2.7080266
		7	0	25.3481	102	0.2036009	0.1042573
			0.3	23.9257	86	0.2117859	0.1226671
	LASSO	3	0	-17.9146	114	-0.1225102	-0.1958190
			0.3	3.2801	74	0.0421035	-0.1180722
		5	0	19.5945	110	0.1595301	0.0603898
			0.3	20.0520	89	0.1776603	0.0747069
		7	0	1.8168	74	-0.0242922	-0.1168795
			0.3	3.5199	66	0.0481258	-0.1276575
Literature (Table 23)	None	3	0	41.0052	92	0.3647555	0.2543023
			0.3	24.1797	99	0.1969891	0.1000813
		5	0	3.1839	66	0.0227622	-0.1108847
			0.3	5.4558	84	0.0528086	-0.0788057
		7	0	24.6735	96	0.2549562	0.1234270
			0.3	15.8630	84	0.1565041	0.0598267
	SFFS	3	0	-25.7442	94	-0.2258416	-0.3397011
			0.3	-0.8207	60	0.0592871	-0.9790217
		5	0	25.1208	96	0.2581941	0.1699558
			0.3	22.1315	82	0.2281864	0.1447917
		7	0	2.1438	118	-0.4112234	-0.3753212
			0.3	-8.4498	73	-0.3854948	-1.0176167
	TS	3	0	13.3439	97	0.1403865	-0.0104126
			0.3	23.2204	101	0.2099201	0.0990272
		5	0	-12.8620	100	-0.0965009	-0.1810188
			0.3	-14.9783	82	-0.3454633	-0.6642436
		7	0	21.6362	106	0.1904325	0.0784067
			0.3	25.4432	101	0.1980456	0.1064507
	LASSO	3	0	-15.3891	83	-0.2678970	-0.5097560
			0.3	0.0000	0	0.0000000	0.0000000
		5	0	25.7859	97	-0.1525371	-0.1433821
			0.3	17.3476	91	0.1547676	0.0472092
		7	0	-28.9513	92	-0.2296458	-0.3321763
			0.3	0.0000	0	0.0000000	0.0000000
Buy-and-Hold strategy profitability over the out-of-sample period: 13.25879							

Table 39 – Trading profitability and transaction costs of machine learning algorithms for the top 50 assets of SSE 180 Index

Technical Analysis indicators	Feature selection method	Hidden layers	Dropout	Strategy profitability	Number of transactions	TC_0	TC_{BH}
Literature + market (Tables 23 and 24)	None	3	0	35.6638	95	0.3841626	0.2829234
			0.3	39.0711	88	0.4365381	0.3410273
		5	0	-4.8916	122	-0.2552616	-0.6701513
			0.3	2.5210	83	0.0390550	-0.3591370
		7	0	32.5512	86	0.3893776	0.2914716
			0.3	34.9669	89	0.1855161	-0.2500637
	SFFS	3	0	-2.5505	116	-0.0255353	-0.0989577
			0.3	-2.6532	94	-0.0343513	-0.2762063
		5	0	37.1818	86	0.4339078	0.3332228
			0.3	37.4408	95	0.3885276	0.3011276
		7	0	-1.2046	58	-0.0498450	-0.2126345
			0.3	-0.3102	77	-0.0128684	-0.1337480
	TS	3	0	37.2284	81	0.4705147	0.3657402
			0.3	38.9214	89	0.4267857	0.3356638
		5	0	-3.0324	110	-0.0308629	-0.1212120
			0.3	0.0000	0	0.0000000	0.0000000
		7	0	32.4841	86	0.3875390	0.2822118
			0.3	37.9563	93	0.4033439	0.3105081
	LASSO	3	0	-5.4923	108	-0.0915193	-0.2472650
			0.3	0.0000	0	0.0000000	0.0000000
		5	0	32.5362	84	0.3538346	0.1858880
			0.3	37.1518	90	0.4047782	0.3146568
		7	0	-4.4870	103	-0.0410895	-0.1209700
			0.3	-0.5910	91	-0.0162708	-0.1006414
Literature (Table 23)	None	3	0	35.7799	84	0.3985859	0.2526929
			0.3	36.8370	90	0.4015542	0.3110039
		5	0	2.5242	141	0.0171953	-0.0463309
			0.3	0.0000	0	0.0000000	0.0000000
		7	0	39.6499	82	0.4834228	0.3831606
			0.3	38.0610	96	0.3865356	0.3029205
	SFFS	3	0	-3.8389	76	-0.0422650	-0.1778699
			0.3	0.0000	0	0.0000000	0.0000000
		5	0	31.2168	89	0.3721600	0.2775215
			0.3	39.3754	89	0.4322248	0.3377194
		7	0	-2.5622	113	-0.0459913	-0.1641081
			0.3	0.0000	0	0.0000000	0.0000000
	TS	3	0	36.0714	88	0.4155866	0.3142577
			0.3	31.2031	85	0.3155642	0.1067525
		5	0	-8.7033	117	-0.0796327	-0.1517418
			0.3	0.0000	0	0.0000000	0.0000000
		7	0	30.6606	89	0.3504062	0.2434762
			0.3	33.2796	89	0.3827255	0.2793077
	LASSO	3	0	-1.2243	112	-0.0179711	-0.1001453
			0.3	0.0000	0	0.0000000	0.0000000
		5	0	38.2578	82	0.4765427	0.3721895
			0.3	37.5638	93	0.3970523	0.3105506
		7	0	-1.7949	128	-0.0134817	-0.0870844
			0.3	0.0000	0	0.0000000	0.0000000
Buy-and-Hold strategy profitability over the out-of-sample period: 8.71314							

Table 40 – Trading profitability and transaction costs of machine learning algorithms for assets of Bovespa Index

While the predictive performance was "split" into the big cases of 50% accuracy and 65% accuracy, the actual profitability of the machine learning bases strategies were more homogeneous: basically the profitability oscillated between the value of the Buy-and-Hold strategy, albeit with a big variance, and consistently negatively concentrated – *i.e.:* the models did not manage to yield profits much larger than the average market

level, while they did manage to register fairly high levels of loss, especially for the British and the German markets. Even some cases with 65% out-of-sample accuracy ended up with non-profitable strategies.

Especially when considering the existence of transaction costs, the profitability of the strategies become even less desirable: in many cases a small profit is attainable using a big number of operations, thus demanding the transaction costs TC_0 and TC_{BH} to be proportionally smaller for the strategy to become actually worth executing to generate some gain. Besides, many strategies had negative profitability to start with, such that the transaction cost would also have to be negative for those strategies to be worthwhile. Therefore, on average, the economic gains of the strategies yielded from machine learning techniques tested in this paper were statistically close to zero, reinforcing the implications of the Efficient Market Hypothesis.

The profits of the strategies were especially bad for the British market, where even the Buy-and-Hold gain was negative, possibly due to the period of relative political and economical instability forthcoming the events of the Brexit referendum in the recent years. For the cases in which the algorithm predicted only one class (the "only rise" and "only drop" cases), the profitability were simply zero (in this case the investor never entered the market, with the number of transactions equal to zero) or a single negative value (in this case the investor only bought the asset on day one and predicted that this price would go up all the way to the last day, in which he would still be expecting a price boost, so this investor only operated once, which was buying the asset on the first day).

Again, the effect of the number of hidden layers is not clear; in terms of profitability the same pattern of similarity between the cases with 3 and 7 hidden layers persist; the results differs from the reports of Nakano, Takahashi and Takahashi (2018), in which deeper neural networks yielded strategies with better profitability. The effect of dropout also seems heterogeneous across different hyperparameters and markets. Concerning the choice of candidate features, the "None" case (no feature selection method) showed fairly good profitability when considering only technical analysis indicators from the "Literature side", while for the larger indicator set composed by both Literature and Market experiences, the application of feature selection algorithms yielded a slight overall improvement for some cases.

5.6 Conclusion

This paper analyzed the performance of deep neural network algorithms to predict the stock price movement based on technical analysis indicators taken from recent scientific articles and from specialized trading websites. Using daily data from financial assets that compose seven market indexes around the world between 2008 and 2019, we tested

different setting of hyperparameters, namely number of hidden layers in each neural network and dropout rate, we applied three feature selection methods (Sequential Forward Floating Selection, Tournament Screening and LASSO) on the feature set of technical analysis indicators, using their filtered counterparts to be used as explanatory variables in the training process.

The results indicated that the out-of-sample accuracy rate of the prediction converged to two values – besides the 50% value that represents the market efficiency, a "strange attractor" of 65% also was achieved consistently. Nonetheless, when applying the prediction into a real trading experiment, the profitability of the strategies did not manage to significantly outperform the Buy-and-Hold strategy, while showing more consistent losses in markets that presented higher levels of volatility during the testing period.

The findings of this paper can be of potential interest for scholars for future inquiries in similar lines of research, as many technical analysis indicators that were most picked by feature selection methods were not considered by authors in recent applications on stock price prediction, being used instead by investors on their real-world trading. In this sense, some indicators composed by the combination of other indicators can be used instead of their constitute counterparts, which can diminish the levels of redundant information taken into account for the models and potentially yield better predictive results and asset allocations. Moreover, the values for the maximum transaction cost levels for an investor to reach some economic gain or to outperform the Buy-and-Hold strategy can be used to analyze the overall attractiveness of different financial markets, with an investor potentially willing to operate in markets in which the transaction costs are lower than the thresholds found in this paper.

The combinations of hyperparameters considered in this paper are not exhaustive, as many improvements and additional cases could be executed. For instance, the only activation function that we applied was the Sigmoid function, while there are many other candidate functions such as the Hyperbolic Tangent and the ReLU (Rectified Linear Unit), both very popular in neural network applications. As discussed in Yaohao and Albuquerque (2019), the choice of the function that define the structure of non-linear interactions of the data have a decisive impact on the results, such that we recommend further investigations about the implications of different activation functions in the application of this paper. Other potential improvements include using more training epochs and testing for more values of number of hidden layers other than 3, 5, and 7, as well as testing other cases for the dropout rate of the networks and other feature selection methods, and further analyze the sensibility of the models to alterations in those hyperparameters. Finally, replications of this study considering other time periods and financial assets are also potential future developments, as well as the use of rolling windows to re-calibrate the models with a larger periodicity and to further investigate whether the strategies' prof-

itability can be better in a smaller period, and in which extent the gains can be higher when using a dynamic model to incorporate sudden changes over the historic pattern.

APPENDIX A – Source codes in R

A.1 Application 1

```
dados<-read.csv(file.choose())

library(data.table)
library(dplyr)
library(lubridate)
library(rugarch)
library(tseries)
library(kernlab)
library(foreach)
library(doParallel)
library(doSNOW)
library(forecast)
library(MCS)
library(highfrequency)

### LOAD WORKSPACE
load("bitcoin lowfreq 1.RData")

log(dados[4])
prices<-(dados[4])^T
head(prices)
n <- length(prices)
logret <- log(prices[-1]/prices[-n]) ## training periods
logret[3]
prices[3:4]
prov<-lag(logret)
logretlag<-c()

for (i in 1:length(logret))
{
  logretlag[i+1]<-logret[i]
```

```
}
logretlag<-logretlag[-length(logretlag)]
logretreal<-logret[-1]
logretlagreal<-logretlag[-1]

mediaols<-lm(logretreal~logretlagreal) # fit mean equation by
   OLS

h_proxy<-(logret-mean(logret))^2 ## proxy volatility
h_proxy[-c(1,2)] ## remove initial NAs

volatols<-lm(h_proxy[-c(1,2)]~a_t[-length(a_t)]+h_proxy[-c(1,
   length(h_proxy))]) # variance equation by OLS
fitted(volatols)
residvolatold<-resid(volatols)

rmse<-function(a){sqrt(mean(a^2))} # RMSE
mae<-function(a){mean(abs(a))} # MAE
nmse<-function(a){(mean(a^2))/(var(a)*length(a))} # NMSE

rmse(resid(volatols))
mae(resid(volatols))
nmse(resid(volatols))

# GARCH(1,1) benchmark

gspec.ru <- ugarchspec(mean.model=list(armaOrder=c(0,0)),
   distribution="sstd") # distribution changes to 'norm', 'std'
   and 'sstd'
gfit.ru <- ugarchfit(gspec.ru, logretreal)
coef(gfit.ru)
residuals(gfit.ru)
plot(gfit.ru@fit$sigma, type='l')

# error metrics
rmse(gfit.ru@fit$sigma-h_proxy[-c(1)])
mae(gfit.ru@fit$sigma-h_proxy[-c(1)])
nmse(gfit.ru@fit$sigma-h_proxy[-c(1)])
```

```
provv<-gfit.ru@fit$sigma-h_proxy[-c(1)]
(sum(provv^2)/length(provv))/(sum((gfit.ru@fit$sigma-mean(
    h_proxy[-c(1)]))^2)/(length(provv)-1))

###### SVR-GARCH ######

Kernelgauss<-function(x,y) #Gaussian Kernel
{
  res<-exp(-sigma*(sum((x-y)^2)))
  return(res)
}
class(Kernelgauss) <- "kernel"

base<-matrix(nrow=length(logretreal), ncol=3)
for(i in 1:length(logretreal))
{
  base[i,1]=i
  base[i,2]=logretreal[i]
  base[i,3]=logretlagreal[i]
}
head(base)

zz<-read.csv(file.choose()) ## validation set
head(zz)
linesvalida<-zz[,1]
base[linesvalida[3]]
base[linesvalida[3410],]

tail(linesvalida)

basevalida<-matrix(nrow=length(linesvalida), ncol=3)
for (i in 1:length(linesvalida))
{
  basevalida[i,]<-base[linesvalida[i],]
}
tail(basevalida)

zzz<-read.csv(file.choose()) ## test set
head(zzz)
```

```
linestesta<-zzz[,1]
base[linestesta[3]]
base[linestesta[3410],]

tail(linestesta)

basetesta<-matrix(nrow=length(linestesta), ncol=3)
for (i in 1:length(linestesta))
{
  basetesta[i,]<-base[linestesta[i],]
}
tail(basetesta)

zzzz<-read.csv(file.choose()) ## training set
head(zzzz)
linestreina<-zzzz[,1]
base[linestreina[3]]
base[linestreina[3410],]

tail(linestreina)

basetreina<-matrix(nrow=length(linestreina), ncol=3)
for (i in 1:length(linestreina))
{
  basetreina[i,]<-base[linestreina[i],]
}
tail(basetreina)

tail(basetreina[-c(8486:8504),])

# hyperparameter grid
epsilonpol<-seq(0.05,1,by=0.05)
Cpol<-seq(0.5,5,by=0.5)
sigmapol<-seq(0.05,2,by=0.05)
parametropol<-as.data.frame(expand.grid(epsilon=epsilonpol,C=
   Cpol,sigma=sigmapol)) # all combinations
performacepol<-as.data.frame(rep(0.0,nrow(parametropol))) #
   receives RMSEs from parameters applied to validation set
```

```
## SVR-GARCH mean equation

acuracia <-foreach(i=(1:nrow(parametropol)),.packages="kernlab",
    .combine='rbind') %dopar%
{
  epsilon<-parametropol[i,1]
  C<-parametropol[i,2]
  sigma<-parametropol[i,3]
  Kernelgauss<-function(x,y)
  {
    res<-exp(-sigma*(sum((x-y)^2)))
    return(res)
  }
  class(Kernelgauss) <- "kernel"

  svm <- ksvm(basetreina[,3],basetreina[,2],epsilon=epsilon,C=C,
      kernel="rbfdot",scaled=T)

  # predicted values for each combination applied to validation
      set
  ypred<-predict(svm,validacao[,2])
  CXP<-as.data.frame(as.numeric(ypred))
  # RMSE of prediction in validation set
  CXP$observado<-basevalida[,2]
  CXP$desvio<-CXP[,1]-CXP[,2]
  CXP$desvioquad<-(CXP[,1]-CXP[,2])^2

  # keep each RMSE
  performacepol[i,1]<-sqrt(mean(CXP$desvioquad))
  print(performacepol)
}

which(performacepol==min(performacepol)) # pick smaller RMSE
performacepol[which(performacepol==min(performacepol)),1] # its
    value
parametropol[which(performacepol==min(performacepol)),] # and
    its parameters

## fit mean equation with best parameters
```

```
bestmediasvr <- ksvm(base[,3],base[,2],epsilon=0.85,C=0.5,kpar=
    list(sigma=0.645497224367902),kernel="rbfdot",scaled=T)
residmediasvr<-(predict(bestmediasvr,base[,2])-base[,2])

## volatility equation:

h_svr<-h_proxy[-1]
h_svr_lag<-h_svr[-1]

for (i in 1:length(h_svr))
{
  h_svr_lag[i+1]<-h_svr[i]
}
h_svr_lag<-h_svr_lag[-length(h_svr_lag)]
h_svr_real<-h_svr[-1] ## after losing 2 periods
h_svr_lag_real<-h_svr_lag[-1]

residmediasvr_lag<-residmediasvr
for (i in 1:length(residmediasvr))
{
  residmediasvr_lag[i+1]<-residmediasvr[i]
}
residmediasvr_lag_real<-residmediasvr_lag[-1]

base2<-matrix(nrow=length(residmediasvr), ncol=4)
for(i in 1:length(residmediasvr))
{
  base2[i,1]=i
  base2[i,2]=residmediasvr_lag_real[i]
  base2[i,3]=h_svr_lag_real[i]
  base2[i,4]=h_svr_real[i]
}
head(base2)
base2<-base2[-719,]
tail(base2)

performacepol2<-as.data.frame(rep(0.0,nrow(parametropol))) #
    receives RMSEs from parameters applied to validation set
```

```
acuracia2 <-foreach(i=(1:nrow(parametropol)),.packages="kernlab
    ", .combine='rbind') %dopar%
{
  epsilon<-parametropol[i,1]
  C<-parametropol[i,2]
  cc<-parametropol[i,3]
  Kernelgauss<-function(x,y)
  {
    res<-exp(-sigma*(sum((x-y)^2)))
    return(res)
  }
  class(Kernelgauss) <- "kernel"

  # analogous to mean equation
  svm <- ksvm(c(treinamento2[,2],treinamento2[,3]),treinamento2
      [,4],epsilon=epsilon,C=C,kernel="rbfdot",scaled=T)

  ypred<-predict(svm,validacao2[,4])
  CXP<-as.data.frame(as.numeric(ypred))
  CXP$observado<-validacao2[,4]
  CXP$desvio<-CXP[,1]-CXP[,2]
  CXP$desvioquad<-(CXP[,1]-CXP[,2])^2

  performacepol2[i,1]<-sqrt(mean(CXP$desvioquad))
  print(performacepol2)
}

which(performacepol2==min(performacepol2)) # analogous to mean
    equation
performacepol2[which(performacepol2==min(performacepol2)),1]
parametropol[which(performacepol2==min(performacepol2)),]

## fit volatility equation with best parameters
bestvolatsvr <- ksvm(c(teste2[,2],teste2[,3]),teste2[,4],epsilon
    =0.1,C=5,kpar=list(sigma=0.30151134457776),kernel="rbfdot",
    scaled=T)
residvolatsvr<-(predict(bestmediasvr,teste2[,4])-teste2[,4])

# RESULTS GARCH
```

```
rmse(gfit.ru@fit$sigma-h_proxy[-c(1)])
mae(gfit.ru@fit$sigma-h_proxy[-c(1)])
nmse(gfit.ru@fit$sigma-h_proxy[-c(1)])

# RESULTS SVR-GARCH
rmse(residvolatsvr)
mae(residvolatsvr)
nmse(residvolatsvr)

## DIEBOLD-MARIANO TEST
residodoc<-as.vector(residuals(gfit.ru))
residodorya<-gfit.ru@fit$sigma-h_proxy[-c(1)]

dm.test(residodorya, residvolatsvr, h=1, power=2) # two-tailed

## MODEL CONFIDENCE SET

SSM_b_1 <- MCSprocedure(Loss = btc_res_1, alpha = 0.05, B =
    10000, statistic = "Tmax", k = 2)

#### Procedure is analogous for all training-validation-test
    windows using high frequency data.
```

A.2 Application 2

```
library(data.table)
library(dplyr)
library(bit64)
library(stringr)
library(readxl)
library(reshape2)
library(knitr)
library(ggplot2)
library(plyr)
library(gplots)
library(lubridate)
library(tcltk)
library(googlesheets)
library(gsheet)
library(corrplot)
```

```
library(xtable)
library(tidyverse)
library(magrittr)
library(RnavGraphImageData)
library(dplyr)
library(e1071)
library(kernlab)
library(quadprog)
library(PerformanceAnalytics)

############ IMPORTING DATA
rm(list = ls());gc()
arquivo1 <- "data/BOLSAS2.xlsx"
planilhas1 <- readxl::excel_sheets(arquivo1)

arquivo2 <- "data/BOLSAS2 - Copy.xlsx"
planilhas2 <- readxl::excel_sheets(arquivo2)


arquivo <- bind_rows(data.frame(planilha = planilhas1,arquivo =
   arquivo1),
                     data.frame(planilha = planilhas2,arquivo =
                        arquivo2))

arquivo <- arquivo %>%
  filter( !grepl("sheet|label|indices",planilha,ignore.case = T)
      ) %>%
  mutate(indice = ifelse(planilha=="Brasil", "IBOV",
  ifelse(planilha=="China_shanghai", "SZSMEC",
  ifelse(planilha=="Franca", "CAC",
  ifelse(planilha=="Alemanha", "SPX",
  ifelse(planilha=="Japao", "NKY",
  ifelse(planilha=="Holanda", "AEX",
  ifelse(planilha=="UK", "UKX","NDX"))))))))



base_indice <- read_excel("data/BOLSAS2 - Copy.xlsx",
                          sheet = "INDICES")
```

```
names(base_indice)[1] <- "DT"

detach(package:plyr)
j <- 2
for (j in (1:nrow(arquivo))[-c(1,3,5)] ) {
  i <- arquivo$planilha[j]
  base <- read_excel(arquivo$arquivo[j],
                     sheet = i)

  names(base)[1] <- "DT"
  names(base) <- gsub("(.*)\\s.*\\sEquity","\\1",names(base))


  base_indice_temp <- base_indice

  names(base_indice_temp)[grep(arquivo$indice[j],
                               names(base_indice_temp))] <- "r"

  base_indice_temp <- base_indice_temp %>% select(DT,r)

  base <- base %>%
    left_join(base_indice_temp,
              by = "DT")


  nomes_colunas_transform <- setdiff(names(base),"DT")
  setDT(base)[, (nomes_colunas_transform):= lapply(.SD,
                                                   function(x)
                                                     as.numeric
                                                     (gsub
                                                     (",",".",x
                                                     ))),
              .SDcols=nomes_colunas_transform]




  ### risk free asset
  base_rf <- readRDS("data/rf.rds")
```

```
base_rf <- base_rf %>%
  mutate(DT = as.POSIXct(DT, format="%Y-%m-%d") ,
         RFe = treasury_10) %>%
  select(DT,RFe)


base <- base %>%
  left_join(base_rf,
            by = "DT") %>% as.data.table()

base[,RF:=ifelse(is.na(RFe),
                 lag(RFe),
                 RFe)]

base <- base %>%
  select(-RFe)

### correct non-existant dates


base <- base  %>%
  gather(key = "var",value = "valor",-DT) %>% data.table()

base <- base[order(var,DT)]
base[,ra:=log(valor)-lag(log(valor)),by="var"]
base[,m:=mean(ra,na.rm=T),by = "var"]
base[,v:=sd(ra,na.rm=T),by = "var"]
base[,value_z:=(ra-m)/v]


base <-base %>% filter(!is.na(ra)) %>%
  select(-value_z,-valor,-m,-v) %>%
  spread(key = var,value = ra) %>%
  filter(!is.na(DT))      %>%
  as.data.table()


dados_faltantes <- t(base[, lapply(.SD, function(x) sum(is.na(
  x)))])
```

```
  dados_faltantes <- cbind(data.table(dados_faltantes),
                           rownames(dados_faltantes)) %>% data.
                              table()
  dados_faltantes <- dados_faltantes[order(V1)]

  ## less than 10% of missing-values
  vars_sem_na <- dados_faltantes$V2[dados_faltantes$V1<(nrow(
     base)/10)]
  length(vars_sem_na)

  saveRDS(base,paste0("data//global//",gsub("\\s","_NOVO_",i),".
     rds"))
  rm(base)
}


############

## Markowitz efficient frontier; no short-selling allowed
eff.frontier <- function (covariance,vec_mean, short="no", max.
   allocation=NULL,
                          risk.premium.up=.5, risk.increment
                             =.005){
  n <- ncol(covariance)
  # Equality constraint
  Amat <- matrix (1, nrow=n)
  bvec <- 1
  meq <- 1

  # Update Amat and bvec
  if(short=="no"){
    Amat <- cbind(1, diag(n))
    bvec <- c(bvec, rep(0, n))
  }

  # Calculate number of loops
  loops <- risk.premium.up / risk.increment + 1
  loop <- 1
```

```
  eff <- matrix(nrow=loops, ncol=n+3)
  colnames(eff) <- c(names(vec_mean), "Std.Dev", "Exp.Return", "
     sharpe")
  i <- 0
  # Solving quadratic optimization problem, gets portfolio with
     greatest Sharpe ratio
  for (i in seq(from=0, to=risk.premium.up, by=risk.increment)){
    dvec <- vec_mean * i
    sol <- solve.QP(covariance, dvec=dvec, Amat=Amat, bvec=bvec,
        meq=meq)
    eff[loop,"Std.Dev"] <- sqrt(sum(sol$solution*colSums((
       covariance*sol$solution))))
    eff[loop,"Exp.Return"] <- as.numeric(sol$solution %*%
       vec_mean)
    eff[loop,"sharpe"] <- eff[loop,"Exp.Return"] / eff[loop,"Std
       .Dev"]
    eff[loop,1:n] <- sol$solution
    loop <- loop+1
  }

  return(as.data.frame(eff))
}

### Finding optimal portfolio for each different covariance
    matrix

vec_parms <- NULL
arquivos <- list.files(path = "data/global",full.names = T)
nomes_arquivos <- gsub(".rds","",list.files(path = "data/global
    "),fixed = T)
arquivos <- list.files(path = "C:/Users/b05652877465/Desktop/
    KERNEL PCA/ROBUST/dados",full.names = T)
nomes_arquivos <- gsub(".rds","",list.files(path = "C:/Users/
    b05652877465/Desktop/KERNEL PCA/ROBUST/dados"),fixed = T)
i <- 4

n_in_sample <- 100/15
```

```
for (i in 1:length(nomes_arquivos)) {
  base <- readRDS(arquivos[i])
  data_in_sample <- rev(base$DT)[trunc(nrow(base)/n_in_sample)]
  var_modelo <- setdiff(names(base),"DT")
  dados_in_sample <- base %>%
    filter(DT<=data_in_sample)

  dados_out_of_sample <- base %>%
    filter(DT>data_in_sample)

  RF_mean <- mean(dados_out_of_sample$RF,na.rm=T)

  ### PEARSON ----

  mpearson  <- cor(dados_in_sample %>% select(var_modelo))
  covariance <- var(dados_in_sample %>% select(var_modelo))
  vec_mean <- colMeans(dados_in_sample %>% select(var_modelo))



  # Run efficient frontier
  eff <- eff.frontier(covariance,vec_mean, short="no", max.
      allocation=NULL,
                      risk.premium.up=1, risk.increment=.1)

  # Optimal portfolio = largest Sharpe ratio
  eff.optimal.point <- eff %>% filter(sharpe==max(eff$sharpe))

  # Checking results
  wm <- eff.optimal.point %>% select(var_modelo)

  A <-  as.matrix(data.matrix(dados_out_of_sample %>% select(
      var_modelo)),
                  ncol = ncol(dados_out_of_sample)) %*%  t(as.
                     matrix(round(wm,10)))

  dados_out_of_sample <- dados_out_of_sample %>%
    mutate(rs = cumsum(r))
  q <- 1
```

```
### Risk Free
RF_mean <- mean(dados_out_of_sample$RF, na.rm=T)


# Market
auto_cov_sp <- acf(dados_out_of_sample$r, lag=1)$acf[2]
n_q_sp <- q^(1/2)*(1 + (2*auto_cov_sp)/(1- auto_cov_sp)*(1 -
    (1-auto_cov_sp^q)/(q*(1-auto_cov_sp))))^(-1/2)

sr_sp <- (mean(dados_out_of_sample$rs)-RF_mean)/sd(
    dados_out_of_sample$rs)
sr_sp_month <-  n_q_sp*sr_sp

# Portfolio
auto_cov_out <- acf(A, lag=1)$acf[2]
n_q <- q^(1/2)*(1 + (2*auto_cov_out)/(1- auto_cov_out)*(1 -
    (1-auto_cov_out^q)/(q*(1-auto_cov_out))))^(-1/2)
sr <- (mean(A)-RF_mean)/sd(A)


# Sharpe ratio
sr_month <- n_q*sr

sr_month_0 <- sr_month
VIID_q <- (n_q^2)*(1+1/2*(sr^2))
VGMM_add <- ifelse(q==1,0,q*(sr)^2*(sum((1-  (1:(q-1))/q)^2)))
VGMM_q <- VIID_q    + VGMM_add

sd_sr_q <- (VGMM_q/nrow(dados_out_of_sample))^(1/2)

# Sortino ratio


sor <- n_q*(mean(A)-RF_mean)/DownsideDeviation(A,MAR = RF_mean
    )
sor_sp <- (mean(dados_out_of_sample$rs)-RF_mean)/
    DownsideDeviation(dados_out_of_sample$rs,MAR = RF_mean)
```

```
VIID_q2 <- (n_q^2)*(1+1/2*(sor^2))
VGMM_add2 <- ifelse(q==1,0,q*(sor)^2*(sum((1- (1:(q-1))/q)^2))
    )
VGMM_q2 <- VIID_q2    + VGMM_add2

sd_sr_q2 <- (VGMM_q2/nrow(dados_out_of_sample))^(1/2)


# Saving results
vec_parms_temp <- data.frame(data = nomes_arquivos[i],
      cov_method = "Pearson",
      initial_in = min(dados_in_sample$DT),
      end_in = max(dados_in_sample$DT),
      initial_out = min(dados_out_of_sample$DT),
      end_out = max(dados_out_of_sample$DT),
      N_in = nrow(dados_in_sample),
      N_out = nrow(dados_out_of_sample),
      X = length(var_modelo),
      X_pos_n = length(which(wm>0.00)),
      X_pos_valid_n = length(which(wm>0.01)),
      X_pos_valid_sum = paste0(round(sum(wm[(which(wm>0.01))])
          *100,2),"%"),
      w_max = paste0(round(max(wm[(which(wm>0.01))])*100,2)
          ,"%"),
      w_min = paste0(round(min(wm[(which(wm>0.01))])*100,2)
          ,"%"),
      r_final = paste0(round(cumsum(A)[nrow(dados_out_of_sample
          )]*100,4),"%"),
      Q = NA,
      lamda_max = NA,
      eigen_up = NA,
      eigen_v_up = NA,
      eigen_max = NA,
      eigen_v_max = NA,
      st_dev=sd(A),
      down_dev=DownsideDeviation(A,MAR = RF_mean),
      sharpe_ratio = sr_month ,
      p_sharp_ratio_sp = 1-pnorm(sr_month-sr_sp_month,0,sd_sr_q
          ),
```

```
      p_sharp_ratio_0 = 1-pnorm(sr_month,0,sd_sr_q),
      rmt_improvement = NA ,
      sortino_ratio = sor ,
      p_sort_ratio_sp = 1-pnorm(sor-sor_sp,0,sd_sr_q2),
      p_sort_ratio_0 = 1-pnorm(sor,0,sd_sr_q2),
      auto_cov_ar1 = auto_cov_out,
      p_sharp_ratio_rmt = NA,
      p_sort_ratio_rmt = NA)

vec_parms <- rbind(vec_parms,vec_parms_temp)

### PEARSON RMT ----
Q <- nrow(dados_in_sample)/(length(var_modelo))

## Singular value decomposition of covariance matrix
eigen0 <- eigen(mpearson)
lamda_max <- (1 + 1/Q + (1/Q)^.5)


g1 <- eigen0$values[eigen0$values>=lamda_max]
g2 <- eigen0$values[eigen0$values<lamda_max]
eigen1<- diag(c(g1,rep(mean(eigen0$values),length(g2))))
cor1 <- eigen0$vectors %*% eigen1 %*% t(eigen0$vectors)
covariance_rmt <-  cor1

# Convert to correlation
for(ii in 1:ncol(cor1)){
  for(jj in 1:nrow(cor1)){
    covariance_rmt[ii,jj] <- cor1[ii,jj] * (covariance[ii,ii]*
       covariance[jj,jj])^.5
  }
}

# Efficient frontier
eff <- eff.frontier(covariance_rmt,vec_mean, short="no", max.
   allocation=NULL,
                    risk.premium.up=1, risk.increment=.1)

# Optimal portfolio
```

```
eff.optimal.point <- eff %>% filter(sharpe==max(eff$sharpe))

# Checking
wm <- eff.optimal.point %>% select(var_modelo)

B <-  as.matrix(data.matrix(dados_out_of_sample %>% select(
    var_modelo)),
                  ncol = ncol(dados_out_of_sample)) %*%
   t(as.matrix(round(wm,10)))

# Portfolio
auto_cov_out <- acf(B,lag=1)$acf[2]
n_q <- q^(1/2)*(1 + (2*auto_cov_out)/(1- auto_cov_out)*(1 -
    (1-auto_cov_out^q)/(q*(1-auto_cov_out))))^(-1/2)

# Sharpe ratio
sr <- (mean(B)-RF_mean)/sd(B)
sr_month <- n_q*sr
VIID_q <- (n_q^2)*(1+1/2*(sr^2))
VGMM_add <- ifelse(q==1,0,q*(sr)^2*(sum((1-  (1:(q-1))/q)^2)))
VGMM_q <- VIID_q    + VGMM_add

sd_sr_q <- (VGMM_q/nrow(dados_out_of_sample))^(1/2)

# Sortino ratio
sor <- n_q*(mean(B)-RF_mean)/DownsideDeviation(B,MAR = RF_mean
    )
sor_sp <- (mean(dados_out_of_sample$rs)-RF_mean)/
    DownsideDeviation(dados_out_of_sample$rs,MAR = RF_mean)



VIID_q2 <- (n_q^2)*(1+1/2*(sor^2))
VGMM_add2 <- ifelse(q==1,0,q*(sor)^2*(sum((1-  (1:(q-1))/q)^2))
    )
VGMM_q2 <- VIID_q2    + VGMM_add2

sd_sr_q2 <- (VGMM_q2/nrow(dados_out_of_sample))^(1/2)
```

```
vec_parms_temp <- data.frame(data = nomes_arquivos[i],
    cov_method = "Pearson RMT",
    initial_in = min(dados_in_sample$DT),
    end_in = max(dados_in_sample$DT),
    initial_out = min(dados_out_of_sample$DT),
    end_out = max(dados_out_of_sample$DT),
    N_in = nrow(dados_in_sample),
    N_out = nrow(dados_out_of_sample),
    X = length(var_modelo),
    X_pos_n = length(which(wm>0.00)),
    X_pos_valid_n = length(which(wm>0.01)),
    X_pos_valid_sum = paste0(round(sum(wm[(which(wm>0.01))])
        *100,2),"%"),
    w_max = paste0(round(max(wm[(which(wm>0.01))])*100,2),"%"),
    w_min = paste0(round(min(wm[(which(wm>0.01))])*100,2),"%"),
    r_final = paste0(round(cumsum(B)[nrow(dados_out_of_sample)
        ]*100,4),"%"),
    Q = Q,
    lamda_max = lamda_max,
    eigen_up = paste0(length(g1)," (",round(length(g1)/(ncol(
        dados_in_sample) - 1)*100,2),"%)"),
    eigen_v_up = paste0(round(sum(g1)/(sum(g1)+sum(g2))*100,2)
        ,"%"),
    eigen_max = g1[1]/lamda_max,
    eigen_v_max = paste0(round(sum(g1[1])/(sum(g1)+sum(g2))
        *100,2),"%"),
    st_dev=sd(B),
    down_dev=DownsideDeviation(B,MAR = RF_mean),
    sharpe_ratio = sr_month ,
    p_sharp_ratio_sp = 1-pnorm(sr_month-sr_sp_month,0,sd_sr_q),
    p_sharp_ratio_0 = 1-pnorm(sr_month,0,sd_sr_q),
    rmt_improvement =((mean(B-A)+1)^(252)-1)*100,
    sortino_ratio = sor ,
    p_sort_ratio_sp = 1-pnorm(sor-sor_sp,0,sd_sr_q2),
    p_sort_ratio_0 = 1-pnorm(sor,0,sd_sr_q2),
    auto_cov_ar1 = auto_cov_out,
    p_sharp_ratio_rmt =  1-pnorm(sr_month-sr_month_0,0,sd_sr_q)
        ,
    p_sort_ratio_rmt =  1-pnorm(sor-sor_sp,0,sd_sr_q2))
```

```
  vec_parms <- rbind(vec_parms,vec_parms_temp)

    ## saving results
  dados_out_of_sample$pearson <- cumsum(A)
  dados_out_of_sample$RMT <- cumsum(B)

#### Procedure is analogous for robust covariance estimators and
     Kernel covariance matrices, as well as their respective
    counterparts cleaned by RMT
```

A.3 Application 3

```
library(tidyverse)
library(forecast)
library(quantmod)
library(QuantTools)
library(TTR)
library(keras)
library(xtable)

### Data collection

# vector of assets
vec <- c("AAPL","ABBV","ABT","ACN","AGN","AIG","ALL","AMGN","
   AMZN","AXP","BA","BAC","BIIB","BK","BKNG","BLK","BMY","BRK.B
    ","C","CAT","CELG","CHTR","COST","CL","CMCSA","COF","COP","
   CSCO","CVS","CVX","DHR","DIS","DUK","DWDP","EMR","EXC","F","
   FB","FDX","FOX","FOXA","GD","GE","GILD","GM","GOOG","GOOGL","
   GS","HAL","HD","HON","IBM","INTC","JNJ","JPM","KHC","KMI","KO
    ","LLY","LMT","LOW","MA","MCD","MDLZ","MDT","MET","MMM","MO
    ","MRK","MS","MSFT","NEE","NFLX","NKE","NVDA","ORCL","OXY","
   PEP","PFE","PG","PM","PYPL","QCOM","RTN","SBUX","SLB","SO","
   SPG","T","TGT","TXN","UNH","UNP","UPS","USB","UTX","V","VZ","
   WBA","WFC","WMT","XOM")

for (esse in 1:length(vec)){
  dados <- getSymbols(as.character(vec[esse]),from =
     '2008-01-01',to = '2019-03-01')
```

```
  QQQ <- get(as.character(vec[esse]))
}

lista <- list()

## TA indicators

for (j in 1:length(vec)){
  QQQ <- get(vec[j])
  QQQ <- QQQ %>% na.omit()

   fechamento <- QQQ[,4]
   close <- QQQ[,4]
   high <- QQQ[,2]
   low <- QQQ[,3]
   open <- QQQ[,1]
   volume <- QQQ[,5]

   volume[volume==0] <- mean(volume)

  ###################################

  YYY <- as.xts(lead(as.vector(sign(diff(close)))),order.by =
      index(close))
  YYY[which(YYY==0)] <- -1
  ###################################
  N <- 10
  AB_UP  <- TTR::SMA( QQQ[,3] * (1 - 4*(QQQ[,2]-QQQ[,3] ) / (QQQ
      [,2] + QQQ[,3]) ),
                      order = N)
  ###################################
  AB_DOWN  <- TTR::SMA( high * (1 + 4*(high-low ) / (high + low)
      ),
                        order = N)
  ###################################
  AD <- TTR::chaikinAD(HLC = QQQ[,2:4], volume = QQQ[,5])
  ###################################
  MFM = ((close - low) - (high - close)) / ((high - low))
  ###################################
```

```
MFV = MFM * volume

ADL <- 0
for ( i in 2:dim(QQQ)[1] ) { ADL[i] <- ADL[i-1] + MFV[i]}

ADL <- as.xts(ADL, order.by = index(close))

###################################
ADX <- TTR::ADX(HLC = QQQ[,2:4], n = 14)[,4]
###################################
CHOSC <- TTR::EMA(ADL, n = 3) - TTR::EMA(ADL, n = 10)
###################################
ADO <- (high - diff(close))/(high - low)
###################################
n_fast = 10
n_slow = 20

APO <- TTR::EMA( close , n = n_fast ) - TTR::EMA( close , n =
    n_slow )
###################################
AR_POS <- TTR::aroon(HL = QQQ[,2:3],n = 25)[,1]
AR_NEG <- TTR::aroon(HL = QQQ[,2:3],n = 25)[,2]
AR_OSC <- TTR::aroon(HL = QQQ[,2:3],n = 25)[,3]
###################################
ATR <- TTR::ATR(HLC = QQQ[,2:4],n = 14,maType = TTR::SMA)[,2]
###################################
ATRP <- ATR/close*100
###################################
# Average Volume
AVOL = TTR::SMA(volume , N)
###################################
BB_LOW <- TTR::BBands(HLC = QQQ[,2:4], n = 20, maType = TTR::
    SMA, sd = 2)[,1]
BB_UP <- TTR::BBands(HLC = QQQ[,2:4], n = 20, maType = TTR::
    SMA, sd = 2)[,3]
BB_BW <- (BB_UP-BB_LOW)/TTR::BBands(HLC = QQQ[,2:4], n = 20,
    maType = TTR::SMA, sd = 2)[,2]*100
###################################
bww <- function(close, n){
```

```
  res <- (close - TTR::SMA(close, n))^2
  res <- TTR::SMA(res,n)
  return(res)
}

BWW <- bww(close, n = 10)
###################################
volat <- function(close, n){
  res <- close
  for (i in (n):length(res)){
    res[i] <- (sd(close[(i-n+1):i]))
  }
  res[1:(n-1)] <- NA
  return(res)
}

VOLAT <- volat(close, n = 10)
###################################
perc_b <- function(close,n){
  res <- close
  for (i in (n):length(res)){
    res[i] <- (sd(close[(i-n+1):i]))
  }

  low <- TTR::SMA(close,n) - 2*volat(TTR::SMA(close,n), n)
  high <- TTR::SMA(close,n) + 2*volat(TTR::SMA(close,n), n)
  res <- (close-low)/(high-low)
  return(res)
}

PERC_B <- perc_b(close, n = 20)
###################################
CCI <- TTR::CCI(HLC = QQQ[,2:4], n = 20, maType = TTR::SMA, c
    = 0.015)
###################################
CMF <- TTR::CMF(HLC = QQQ[,2:4], volume = QQQ[,5], n = 20)
###################################
cvol <- function(high, low, n){
  prov <- TTR::EMA(high-low, n)
```

```
  res <- high
  for (i in 11:length(res)){
    res[i] <- (as.numeric(prov[i]) - as.numeric(prov[i-n]))/as
       .numeric(prov[i-10])
  }
  res[1:10] <- NA
  return(res)
}

CVOL <- cvol(high, low, n = 5)
###################################
CMO <- TTR::CMO(close, n = 20)
###################################
maxx <- function(close, n){
  res <- close
  for (i in (n+1):length(res)){
    res[i] <- max(close[(i-n):(i-1)])
  }
  res[1:n] <- NA
  return(res)
}

MAXX <- maxx(close, n = 10)
###################################
minn <- function(close, n){
  res <- close
  for (i in (n+1):length(res)){
    res[i] <- min(close[(i-n):(i-1)])
  }
  res[1:n] <- NA
  return(res)
}

MINN <- minn(close, n = 10)
###################################
CHAND_LONG <- maxx(close,n = 22) - 3*TTR::ATR(HLC = QQQ[,2:4],
    n = 22,maType = TTR::SMA)[,2]
CHAND_SHORT <- maxx(close,n = 22) + 3*TTR::ATR(HLC = QQQ
    [,2:4],n = 22,maType = TTR::SMA)[,2]
```

```
#################################
roc <- function(close, n){
  res <- close
  for (i in (n+1):length(res)){
    res[i] <- (as.numeric(close[i]))/(as.numeric(close[i-n]))
      *100
  }
  res[1:n] <- NA
  return(res)
}

ROC <- roc(close, n = 10)
#################################
COPP <- as.xts(TTR::WMA((roc(close, n = 14) + roc(close, n =
   11)), n = 10),order.by = index(close))
#################################
DPO <- TTR::DPO(close, n = 20, maType = TTR::SMA)
#################################
PDM = diff(high)
NDM = c()
for(i in 1: (length(low) - 1) )  {
  NDM[i] = as.numeric( low[i] ) - as.numeric( low[i+1] )
}

WPDM = c(NA,NA)

for (i in 3:length(PDM) ){
  WPDM[i] = PDM[i-1] - mean(PDM[1:i-1],na.rm=TRUE) + as.
      numeric( PDM[i] )
}

WNDM = c(NA,NA)

for (i in 3:length(PDM) ){
  WNDM[i] = NDM[i-1] - mean(NDM[1:i-1],na.rm=TRUE) + as.
      numeric( NDM[i] )
}

close_low  <- c()
```

```
high_close <- c()
high_low   <- c()

for (i in 2:nrow(QQQ)) {
  high_low       <- high - low
  high_close[i] <- as.numeric( high[i] ) - as.numeric( close[i
     -1] )
  close_low[i]  <- as.numeric( close[i-1] ) - as.numeric( low[
     i] )
}

tr = data.frame(High_menos_Low      =  high_low,
                High_menos_Close_1 =  high_close,
                Close_1_menos_Low  =  close_low )


TR =   apply(tr, 1, max, na.rm = TRUE)

WIR = c(NA,NA)
for (i in 3:length(TR)){
  WIR[i] = TR[i-1] - mean(TR[1:i-1],na.rm=TRUE) + as.numeric(
     TR[i] )
}

PDI = ( WPDM / WIR ) * 100
NDI = ( WNDM / WIR ) * 100
DD = abs(PDI - NDI)
DMI = ( DD / (PDI + NDI)  ) * 100

DMI <- as.xts(DMI, order.by = index(close))
####################################
DONCHIAN <- TTR::DonchianChannel(QQQ[,2:3], n = 10)[,2]
####################################
DEMA = 2 * ( TTR::EMA(close , N) - TTR::EMA( TTR::EMA(close, N
    ), N   )  )
####################################
minimos = c()
maximos = c()
for (i in 1:length(low)) {
```

```
  minimos[i] = min(low[1:i])
  maximos[i] = max(high[1:i])
}

close_menos_min = c()
high_menos_min = c()

for ( i in 1:length(low)){
  close_menos_min[i] = close[i] - minimos[i]
  high_menos_min[i]  = maximos[i] - minimos[i]
}

DSS = as.xts((TTR::EMA( TTR::EMA( close_menos_min) ) / TTR::
   EMA( TTR::EMA( high_menos_min) ) ) * 100
              ,order.by = index(close))
####################################
hl_tm1 = NA
for (i in 2:length(high)) hl_tm1[i] = high[i-1] + low[i-1]

prov = ((high - low)/2 - hl_tm1/2 ) / ((volume/100000000) /
   (high-low))
prov[which(prov %>% is.na())] <- 0
prov[which(prov %>% is.infinite())] <- max(volume)

EMV = TTR::SMA(prov,n = 14)
####################################
EMA <- TTR::EMA(close, n = 10)
####################################
FORCE = TTR::EMA(diff(close)*volume, n=13)
####################################
hull <- function(close, n){
  prov1 <- TTR::WMA(close, n = round(n/2))
  prov2 <- TTR::WMA(close, n = n)
  res <- as.xts((TTR::EMA( TTR::EMA( close_menos_min) ) / TTR
     ::EMA( TTR::EMA( high_menos_min) ) ) * 100
                ,order.by = index(close))
  return(res)
}
HULL <- hull(close, n = 20)
```

```
#################################
n1 = 10 ; nf = 2 ; ns = 30

change <- abs(diff(close, lag = 10))
vovovo <- diff(close)
ER <- close
for (i in 11:length(ER)){
  ER[i] <- as.numeric(change[i])/(sum(abs(vovovo[(i-9):i])))
}

SC <- (ER*(2/(nf+1) - 2/(ns+1)) + (2/(ns+1)))^2

KAMA <- TTR::SMA(close, n = 10)
for (i in 11:length(KAMA)){
  KAMA[i] <- as.numeric(KAMA[i-1]) + SC[i]*(as.numeric(close[i
      ]) - as.numeric(KAMA[i-1]))
}
#################################
KC_M = TTR::EMA( high + low + close / 3 , n = 20)

KC_L = KC_M - 2* (TTR::ATR (QQQ[,2:4], n = 10)$atr )
KC_U = KC_M + 2* (TTR::ATR (QQQ[,2:4], n = 10)$atr )
#################################
mqo <- function(close, n){
   prov <- close
   res1 <- close
   res2 <- close
   res3 <- close
   res4 <- close
   for (i in (n+1):length(prov)){
     base <- prov[(i-n):(i-1)]
     modell <- lm(formula = base~index(base))
     res1[i] <- modell$coefficients[1]
     res2[i] <- modell$coefficients[2]
     res3[i] <- predict(modell, close[i])[1]
     res4[i] <- summary(modell)$coefficients[2,2]
   }
   res1[1:n] <- res2[1:n] <- res3[1:n] <- res4[1:n] <- NA
   res <- cbind(res1, res2, res3, res4)
```

```
  return(res)
}

ppp <- mqo(close, 10)

MQO_ALPHA <- ppp[,1]
MQO_BETA <- ppp[,2]
MQO_PRED <- ppp[,3]
MQO_STD <- ppp[,4]
###################################
MACD <- TTR::MACD(close, nFast = 12, nSlow = 26, maType = TTR
    ::EMA, nSig = 9)[,1]
###################################
MACDH <- TTR::MACD(close, nFast = 12, nSlow = 26, maType = TTR
    ::EMA, nSig = 9)[,2]
###################################
MAE_UP = TTR::SMA(close) + TTR::SMA(close)/4
MAE_LOW = TTR::SMA(close) - TTR::SMA(close)/4
###################################
mass <- function(high, low, n){
  prov <- TTR::EMA(high-low, n) / TTR::EMA(TTR::EMA(high-low,
     n), n)
  res <- high
  for (i in (3*(n-1)+24):length(res)){
    res[i] <- sum(prov[(i-24):i])
  }
  res[1:(3*(n-1)+23)] <- NA
  return(res)
}

MASS <- mass(high, low, n = 9)
###################################
RMF <- (high+low+close)/3*volume
###################################
MFI <- TTR::MFI(HLC = QQQ[,2:4], volume = QQQ[,5], n = 14)
###################################
MIDPOINT <- (MAXX - MINN)/2
###################################
MIDPRICE <- (maxx(high, n = 10) - minn(low, n = 10))/2
```

```
##################################
MOM <- momentum(close, n = 10)
##################################
nvi <- function(close, volume){
  res <- close
  res[1] <- 1000
  for (i in 2:length(res)){
    if (as.numeric(volume[i]) < as.numeric(volume[i-1])){
      res[i] <- as.numeric(res[i-1]) + (as.numeric(close[i]) -
          as.numeric(close[i-1]))/
        as.numeric(close[i-1]) * as.numeric(res[i-1])
    }
    else{
      res[i] <- res[i-1]
    }
  }
  return(res)
}

NVI <- nvi(close, volume)
##################################
NATR <- ATR/close*100
##################################
OBV <- TTR::OBV(close, volume)
##################################
SAR <- TTR::SAR(HL = QQQ[,2:3], accel = c(0.02,0.2))
##################################
TP <- (high + low + close)/3
SS1 <- 2*TP - high
SS2 <- TP - (high - low)
SR1 <- 2*TP - low
SR2 <- TP + (high - low)
##################################
FS1 <- TP - (0.382 * (high - low))
FS2 <- TP - (0.618 * (high - low))
FR1 <- TP + (0.382 * (high - low))
FR2 <- TP + (0.618 * (high - low))
##################################
ddd <- close
```

```
for (i in 1:length(ddd)){
  if (close[i] == open[i]){
    ddd[i] <- high[i] + low[i] + 2*close[i]
  }
  else if (close[i] > open[i]){
    ddd[i] <- 2*high[i] + low[i] + close[i]
  }
  else{
    ddd[i] <- high[i] + 2*low[i] + close[i]
  }
}

PD1 <- ddd/4
DS1 <- (ddd/2) - high
DR1 <- (ddd/2) - low
###################################
pc_up <- function(high, n){
  res <- high
  for (i in n:length(res)){
    res[i] <- max(high[(i-n+1):i])
  }
  res[1:(n-1)] <- NA
  return(res)
}

PC_UP <- pc_up(high, n = 20)
###################################
pc_down <- function(low, n){
  res <- low
  for (i in n:length(res)){
    res[i] <- min(low[(i-n+1):i])
  }
  res[1:(n-1)] <- NA
  return(res)
}

PC_DOWN <- pc_down(low, n = 20)
###################################
chopp <- function(high, low, n){
```

```
  prov1 <- TTR::ATR(HLC = QQQ[,2:4],n,maType = TTR::SMA)[,2]
  prov2 <- pc_up(high, n) - pc_down(low, n)

  res <- close
  res[1:(n-1)] <- NA
  for (i in n:length(res)){
    res[i] <- (log10(sum(prov1[(i-n+1):i])+n))/prov2[i]
  }
  return(res)
}

CHOPPINESS <- chopp(high, low, n = 14)
##################################
ppo <- function(close, nfast, nslow){
  res <- (EMA(close, nfast) - EMA(close, nslow))/EMA(close,
    nslow) * 100
  return(res)
}

PPO <- ppo(close, nfast = 12, nslow = 26)
##################################
PPOH <- PPO - EMA(PPO, n = 9)
##################################
pvo <- function(volume, nfast, nslow){
  res <- (EMA(volume, nfast) - EMA(volume, nslow))/EMA(volume,
    nslow) * 100
  return(res)
}

PVO <- pvo(volume, nfast = 12, nslow = 26)
##################################
PVOH <- PVO - EMA(PVO, n = 9)
##################################
pvi <- function(close, volume){
  res <- close
  res[1] <- 1000
  for (i in 2:length(res)){
    if (as.numeric(volume[i])>as.numeric(volume[i-1])){
      res[i] <- as.numeric(res[i-1]) + (as.numeric(close[i]) -
```

```
            as.numeric(close[i-1]))/
          as.numeric(close[i-1]) * as.numeric(res[i-1])
      }
      else{
        res[i] <- res[i-1]
      }
    }
    return(res)
}

PVI <- pvi(close, volume)
##################################
pvt <- function(close, volume){
    res <- close
    res[1] <- 0
    for (i in 2:length(res)){
      res[i] <- as.numeric(res[i-1]) + ((as.numeric(close[i]) -
          as.numeric(close[i-1]))/
                                          as.numeric(close[i-1])
                                            * as.numeric(
                                          volume[i]))
    }
    return(res)
}

PVT <- pvt(close, volume)
##################################
KST <- TTR::SMA(roc(close, n = 10),n = 10) + 2*(TTR::SMA(roc(
    close, n = 15),n = 10)) +
    3*(TTR::SMA(roc(close, n = 20),n = 10)) + 4*(TTR::SMA(roc(
        close, n = 30),n = 15))
##################################
PSK <- TTR::SMA(roc(close, n = 10),n = 10) + 2*(TTR::SMA(roc(
    close, n = 15),n = 10)) +
    3*(TTR::SMA(roc(close, n = 20),n = 10)) + 4*(TTR::SMA(roc(
        close, n = 30),n = 15)) +
   TTR::SMA(roc(close, n = 40),n = 50) + 2*(TTR::SMA(roc(close,
        n = 65),n = 65)) +
    3*(TTR::SMA(roc(close, n = 75),n = 75)) + 4*(TTR::SMA(roc(
```

```
      close , n = 100),n = 100)) +
  TTR::SMA(roc(close , n = 195),n = 130) + 2*(TTR::SMA(roc(
      close , n = 265),n = 130)) +
  3*(TTR::SMA(roc(close , n = 390),n = 130)) + 4*(TTR::SMA(roc(
      close , n = 530),n = 195))
##################################
RSI <- RSI(close , n = 10, maType = TTR::SMA)
##################################
pp1 <- TTR::SMA(((close-open) + 2*(diff(close)-diff(open)) +
                  2*(diff(close,lag = 2)-diff(open,lag = 2))
                    + (diff(close,lag = 3)-diff(open,lag =
                    3)))/6, n = 12)
pp2 <- TTR::SMA(((high-low) + 2*(diff(high)-diff(low)) +
                  2*(diff(high,lag = 2)-diff(low,lag = 2)) +
                    (diff(high,lag = 3)-diff(low,lag = 3)))
                    /6, n = 12)

RVI <- pp1/pp2
##################################
SMA <- TTR::SMA(close , n = 10)
##################################
stochrsi <- function(close , n){
  prov <- RSI(close , n = 10, maType = TTR::SMA)
  res <- close
  res[1:n] <- NA
  for (i in (n+1):length(res)){
    res[i] <- (prov[i] - min(prov[(i-n):(i-1)]))/(max(prov[(i-
      n):(i-1)]) - min(prov[(i-n):(i-1)]))
  }
  return(res)
}

STRSI <- stochrsi(close , n = 10)
##################################
STOCH_K<- stoch(close , nFastK = 10, nSlowD = 3, nFastD = 3)
    [,1]
##################################
STOCH_D<- stoch(close , nFastK = 10, nSlowD = 3, nFastD = 3)
    [,2]
```

```
##################################
STOCH_D_SLOW<- stoch(close, nFastK = 10, nSlowD = 3, nFastD =
    3)[,3]
##################################
TEMA <- (3*EMA(close, n = 10)) -
   (3*EMA(EMA(close, n = 10), n = 10)) +
   (3*EMA(EMA(EMA(close, n = 10), n = 10), n = 10))
##################################
TRIMA <- SMA(close, n = round((10+1)/2))
##################################
TRIX <- TRIX(close, n = 15, maType = TTR::EMA)[,1]
##################################
TSI <- (EMA(EMA(diff(close),n = 25),n = 13))/(EMA(EMA(abs(diff
    (close)),n = 25),n = 13))
##################################
TP <- (high + low + close)/3
##################################
ulcer <- function(close, n){
   prov <- close
   prov[1:n] <- NA
   for (i in (n+1):length(prov)){
     prov[i] <- (as.numeric(close[i]) - as.numeric(max(close[(i
        -n+1):i])))/
       as.numeric(max(close[(i-n+1):i])) * 100
   }
   res <- sqrt(SMA(prov^2,n))
   return(res)
}

ULCER <- ulcer(close, n = 14)
##################################
ULTOSC <- ultimateOscillator(QQQ[,2:4], n = c(7,14,28), wts =
    c(4,2,1))
##################################
vama <- function(close, volume, n){
   avol <- SMA(volume, n)
   prov <- (3*volume)/(2*avol)*close
   return(SMA(prov, n))
}
```

```
VAMA <- vama(close, volume, n = 10)
#################################
vwap <- function(volume, n){
  propro <- TP*volume
  res <- volume
  res[1:(n-1)] <- NA
  for (i in n:length(res)){
    res[i] <- (sum(propro[(i-n+1):i]))/(sum(volume[(i-n+1):i])
        )
  }
  return(res)
}

VWAP <- vwap(volume, n = 15)
#################################
vol_osc <- function(volume, n_fast, n_slow){
  return((SMA(volume, n_fast) - SMA(volume, n_slow))/SMA(
      volume, n_slow)*100)
}

VOOSC <- vol_osc(volume, n_fast = 14, n_slow = 28)
#################################
vpt <- function(volume, close){
  res <- volume
  res[1] <- 0
  for (i in 2:length(res)){
    res[i] <- as.numeric(res[i-1]) + as.numeric(volume[i])*
      ((as.numeric(close[i]) - as.numeric(close[i-1]))/as.
          numeric(close[i-1]))
  }
  res[1] <- NA
  return(res)
}

VPT <- vpt(volume, close)
#################################
pvoi <- function(high, low, n){
  prov <- abs(high - diff(low))
```

```
  tr <- TTR::ATR(HLC = QQQ[,2:4],n = 14, maType = TTR::SMA)
     [,1]
  res <- high
  for (i in (n+2):length(res)){
    res[i] <- (sum(prov[(i-n):(i-1)]))/(sum(tr[(i-n):(i-1)]))
  }
  res[1:(n+1)] <- NA
  return(res)
}

nvoi <- function(high, low, n){
  prov <- abs(low - diff(high))
  tr <- TTR::ATR(HLC = QQQ[,2:4],n = 14, maType = TTR::SMA)
     [,1]
  res <- high
  for (i in (n+2):length(res)){
    res[i] <- (sum(prov[(i-n):(i-1)]))/(sum(tr[(i-n):(i-1)]))
  }
  res[1:(n+1)] <- NA
  return(res)
}

PVOI <- pvoi(high, low, n = 14)
NVOI <- nvoi(high, low, n = 14)
################################
WILL_R <- WPR(QQQ[,2:4],n = 10)
################################
WMA <- WMA(close, n = 10, wts = 1:10)
################################
wws <- function(close,n){
  res <- close
  res[1:(n-1)] <- NA
  res[n] <- SMA(close, n)[n]
  for (i in (n+1):length(res)){
    res[i] <- as.numeric(res[(i-1)]) + (close[i] - as.numeric(
       res[(i-1)]))/n
  }
  return(res)
}
```

```
WWS <- wws(close, n = 10)
##################################
DISP <- fechamento/(SMA(x = fechamento, n = 9))
##################################
n_fast <- 12
n_slow <- 24

OSCP <- (SMA(x = fechamento, n = n_fast)-SMA(x = fechamento, n
    = n_slow))/SMA(x = fechamento, n = n_fast)
##################################
indic_up <- function(x,n){
  indic <- x
  for (i in 2:length(indic)){
    indic[i] <- sign(as.numeric(x[i])-as.numeric(x[i-1]))
  }
  indic[1] <- NA

  res <- x
  for (i in (n+1):length(res)){
    res[i] <- length(which(indic[(i-n+1):(i)]=="1"))
  }
  res[1:n] <- NA

  return(res)
}

PSY <- function(x,n){
  return((indic_up(x,n))/n*100)
}

PSY <- PSY(fechamento,n = 10)
##################################
DIU <- function(close,high,low,n){

  numm <- close
  for (i in (n+1):length(close)){
    numm[i] <- sum(diff(high)[(i-n+1):i])
  }
```

```
  numm[1:n] <- NA

  prov <- cbind(as.numeric(high - low),
                c(NA,as.numeric(high[-1]) - as.numeric(
                  fechamento[-length(fechamento)])),
                c(NA,as.numeric(as.numeric(fechamento[-length(
                  fechamento)]) - low[-1])))

  prov2 <- as.xts(apply(prov,1,max),order.by = index(close))

  res <- numm/prov2
  return(res*100)
}

DIU <- DIU(close, high, low, n=10)
####################################
DID <- function(close,high,low,n){

  numm <- close
  for (i in (n+1):length(close)){
    numm[i] <- sum(diff(low)[(i-n+1):i])
  }
  numm[1:n] <- NA

  prov <- cbind(as.numeric(high - low),
                c(NA,as.numeric(high[-1]) - as.numeric(
                  fechamento[-length(fechamento)])),
                c(NA,as.numeric(as.numeric(fechamento[-length(
                  fechamento)]) - low[-1])))

  prov2 <- as.xts(apply(prov,1,max),order.by = index(close))

  res <- numm/prov2
  return(res*100)
}

DID <- DID(close, high, low, n=10)
####################################
BIAS <- (close - SMA(close,n=5))/SMA(close,n=5)*100
```

```
##################################
vol_ratio <- function(close,volume,n){
  indic <- close
  for (i in 2:length(indic)){
    indic[i] <- sign(as.numeric(close[i])-as.numeric(close[i
       -1]))
  }
  indic[1] <- NA

  res1 <- vector(mode = "numeric",length = length(close))
  for (i in 2:length(res1)){
    if (indic[i]>0){
      res1[i] <- volume[i]
    }
    else{
      res1[i] <- 0
    }
  }
  res1[1] <- NA

  res2 <- vector(mode = "numeric",length = length(close))
  for (i in 2:length(res2)){
    if (indic[i]<=0){
      res2[i] <- volume[i]
    }
    else{
      res2[i] <- 0
    }
  }
  res2[1] <- NA

  res <- close

  for (i in (n+1):length(close)){
    if (as.numeric(sum(res2[(i-n+1):i]) - sum(volume[(i-n+1):i
       ])) == 0){
      res[i] <- 0
    }
    else{
```

```
        res[i] <- as.numeric((sum(res1[(i-n+1):i]) - sum(volume
            [(i-n+1):i])))/
          as.numeric(sum(res2[(i-n+1):i]) - sum(volume[(i-n+1):i
            ]))
      }
    }
    res[1:n] <- NA

    return(res)
}

VOLR <- vol_ratio(close,volume,n = 3)
##################################
a_ratio <- function(open, high, low, n){
    prov1 <- high-open
    prov2 <- open-low

    res <- low
    for (i in (n):length(res)){
      res[i] <- (sum(prov1[(i-n+1):i]))/(sum(prov2[(i-n+1):i]))
    }
    res[1:(n-1)] <- NA

    return(res)
}

ARATIO <- a_ratio(open, high, low, n = 10)
##################################
b_ratio <- function(close, high, low, n){
    prov1 <- high-close
    prov2 <- close-low

    res <- close
    for (i in (n):length(res)){
      res[i] <- (sum(prov1[(i-n+1):i]))/(sum(prov2[(i-n+1):i]))
    }
    res[1:(n-1)] <- NA

    return(res)
```

```
}

BRATIO <- b_ratio(close, high, low, n = 10)
##################################
REX <- SMA(3*close - (low + open + high), n=20)
##################################
hpr <- function(close,n){
  res <- close
  for (i in n:length(close)){
    res[i] <- res[i]/max(close[(i-n+1):i])
  }
  res[1:(n-1)] <- NA
  return(res)
}

HPR <- hpr(close,n=10)
##################################
lpr <- function(close,n){
  res <- close
  for (i in n:length(close)){
    res[i] <- min(close[(i-n+1):i])/res[i]
  }
  res[1:(n-1)] <- NA
  return(res)
}

LPR <- lpr(close,n=10)
##################################
vmom <- function(volume,n){
  res <- volume
  for (i in (n+1):length(res)){
    res[i] <- as.numeric(volume[i])-as.numeric(volume[i-n])
  }
  res[1:n] <- NA
  return(res)
}

VMOM <- vmom(volume, n = 10)
##################################
```

```
mpp <- function(close, n){
  res <- close
  for (i in n:length(close)){
    res[i] <- (close[i] - min(close[(i-n+1):i]))/(max(close[(i
       -n+1):i]) - min(close[(i-n+1):i]))
  }
  res[1:(n-1)] <- NA
  return(res)
}

MPP <- mpp(close, n = 10)
#################################
var_ratio <- function(close, n){
  res <- close
  for (i in (2*n):length(res)){
    res[i] <- ((sd(close[(i-n+1):i]))^2)/((sd(close[(i-n-n+1)
       :(i-n)]))^2)
  }
  res[1:(2*n-1)] <- NA
  return(res)
}
VARR <- var_ratio(close, n = 10)
#################################
EMA <- TTR::EMA(close, n = 10)
#################################
ATIVO <- j
#################################

BASIS <- cbind(ATIVO, AB_UP,AB_DOWN, AD, MFM, ADL, ADX, CHOSC,
    ADO, APO, AR_POS, AR_NEG, AR_OSC, ATR, ATRP, AVOL, BB_UP,
   BB_LOW, BB_BW, BWW, VOLAT, PERC_B, CCI, CMF, CVOL, CMO,
   MAXX, MINN, CHAND_LONG, CHAND_SHORT, ROC, COPP, DPO, DMI,
   DONCHIAN, DEMA, DSS, EMA, EMV, FORCE, HULL, KAMA, KC_U,
   KC_M, KC_L, MQO_ALPHA, MQO_BETA, MQO_PRED, MQO_STD, MACD,
   MACDH, open, close, volume, MAE_UP, MAE_LOW, MASS, RMF, MFI
    , MIDPOINT, MIDPRICE, MOM, NVI, NATR, OBV, SAR, TP, SS1,
   SS2, SR1, SR2, FS1, FS2, FR1, FR2, PD1, DS1, DR1, PC_UP,
   PC_DOWN, CHOPPINESS, PPO, PPOH, PVO, PVOH, PVI, PVT, KST,
   PSK, RSI, RVI, SMA, STRSI, STOCH_D, STOCH_K, STOCH_D_SLOW,
```

```
    TEMA, TRIMA, TRIX, TSI, ULCER, ULTOSC, VAMA, VWAP, VOOSC,
    VPT, PVOI, NVOI, WILL_R, WMA, WWS, DISP, OSCP, PSY, DIU,
    DID, BIAS, VOLR, ARATIO, BRATIO, REX, HPR, LPR, VMOM, MPP,
    VARR, YYY)

  colnames(BASIS) <- c("ATIVO","AB_UP","AB_DOWN","AD","MFM","ADL
    ","ADX","CHOSC","ADO","APO","AR_POS","AR_NEG","AR_OSC","ATR
    ","ATRP","AVOL","BB_UP","BB_LOW","BB_BW","BWW","VOLAT","
    PERC_B","CCI","CMF","CVOL","CMO","MAXX","MINN","CHAND_LONG
    ","CHAND_SHORT","ROC","COPP","DPO","DMI","DONCHIAN","DEMA
    ","DSS","EMA","EMV","FORCE","HULL","KAMA","KC_U","KC_M","
    KC_L","MQO_ALPHA","MQO_BETA","MQO_PRED","MQO_STD","MACD","
    MACDH","OPEN","CLOSE","VOLUME","MAE_UP","MAE_LOW","MASS","
    RMF","MFI","MIDPOINT","MIDPRICE","MOM","NVI","NATR","OBV","
    SAR","TP","SS1","SS2","SR1","SR2","FS1","FS2","FR1","FR2","
    PD1","DS1","DR1","PC_UP","PC_DOWN","CHOPPINESS","PPO","PPOH
    ","PVO","PVOH","PVI","PVT","KST","PSK","RSI","RVI","SMA","
    STRSI","STOCH_D","STOCH_K","STOCH_D_SLOW","TEMA","TRIMA","
    TRIX","TSI","ULCER","ULTOSC","VAMA","VWAP","VOOSC","VPT","
    PVOI","NVOI","WILL_R","WMA","WWS","DISP","OSCP","PSY","DIU
    ","DID","BIAS","VOLR","ARATIO","BRATIO","REX","HPR","LPR","
    VMOM","MPP","VARR","YYY")

  lista[[j]] <- BASIS %>% na.omit()
  print(j)
  write.csv2(BASIS, file = paste0(vec[j],".csv"))
}

base_completa <- do.call( rbind, lista )

save(base_completa, file = "basecompleta_USA.RData")
write.csv2(base_completa, file = "basecompleta_USA.csv")

### Deep neural networks

# Split data into training and test subsets
col_Y <- which(colnames(base_completa) == "YYY")

treino_id <- seq_len( floor(0.75*nrow(base_completa) ))
```

```
teste_id  <- (length(treino_id) + 1 : nrow(base_completa))

medias  <- apply(X = base_completa[treino_id   , - col_Y], MARGIN
    = 2, mean)
desvios <- apply(X = base_completa[treino_id   , - col_Y], MARGIN
    = 2, sd   )

x_test <- list()
    for (i in 1: length(medias) ) {
      x_test[[i]]  <- (base_completa[-treino_id , - col_Y][,i] -
          medias[i]) / desvios[i]
    }

x_test  <- data.frame(x_test)
names(x_test) <- colnames(base_completa)[-col_Y]

x_train <- base_completa[treino_id   , - col_Y] %>% scale()

colunas_nan1 <- which( apply( rbind(x_train , x_test), 2,
                              function(x) is.nan(x) %>% sum() )
                                > 0 )

x_train <- x_train[ , - which(colnames(x_train)  %in% names(
   colunas_nan1) ) ]
x_test  <- x_test [ , - which(colnames(x_test)   %in% names(
   colunas_nan1) ) ]
x_test <- as.matrix(x_test)

y_train <- ifelse( base_completa[treino_id,col_Y] == 1  , 1  , 0)
   %>%
          as.matrix() %>%
          to_categorical(y_train, num_classes = 2)

y_test  <- ifelse( base_completa[-treino_id,col_Y] == 1  , 1  , 0)
    %>%
          as.matrix() %>%
          to_categorical(y_train, num_classes = 2)

preco_puro <- base_completa[ - treino_id, "CLOSE" ]
```

```
resultados <- list()

colunas <- list()

# Columns picked by feature selection methods
colunas$todas <- colnames(x_train)
colunas$todas <- setdiff(colunas$todas, names(colunas_nan1))

colunas$todas_tab1 <- c("SMA","WMA","EMA","MOM","STOCH_K","
   STOCH_D","STOCH_D_SLOW","RSI","MACD","WILL_R","ADO","CCI","
   ROC","DISP","OSCP","PSY","DIU","DID","BIAS","VOLR","ARATIO","
   BRATIO","ATR","BB_UP","BB_LOW","DMI","KC_U","KC_L","TRIMA","
   MAE_UP","MAE_LOW","REX","NVI","PVI","VAMA","HPR","LPR","MAXX
    ","MINN","VMOM","MPP","PPO","SAR","OBV","VOLAT","MFI","VARR
    ","MQO_BETA","OPEN","CLOSE","VOLUME")
colunas$todas_tab1 <- setdiff(colunas$todas_tab1, names(
   colunas_nan1))

colunas$lasso <- c("MFM","ADL","ADX","CHOSC","ADO","APO","AR_NEG
    ","ATRP","AVOL","BB_BW","BWW","PERC_B","CCI","CMF","CVOL","
   ROC","COPP","DPO","DMI","DEMA","DSS","EMV","FORCE","HULL","
   MQO_ALPHA","MQO_BETA","MQO_STD","CLOSE","VOLUME","RMF","MOM
    ","NVI","NATR","OBV","DS1","DR1","CHOPPINESS","PVO","PVOH","
   PSK","RSI","RVI","STRSI","STOCH_D","STOCH_K","STOCH_D_SLOW","
   TSI","ULTOSC","VAMA","PVOI","PSY","DIU","BIAS","VOLR","BRATIO
    ","REX","HPR","VMOM","MPP","VARR")
colunas$lasso <- setdiff(colunas$lasso,names(colunas_nan1) )

colunas$lasso_tab1 <- c("SMA","WMA","MOM","STOCH_K","STOCH_D","
   STOCH_D_SLOW","RSI","MACD","WILL_R","CCI","ROC","DISP","OSCP
    ","PSY","DIU","DID","BIAS","VOLR","ARATIO","BRATIO","BB_UP","
   DMI","REX","NVI","PVI","HPR","LPR","VMOM","MPP","PPO","SAR","
   OBV","VOLAT","MFI","VARR","MQO_BETA","OPEN","CLOSE","VOLUME")
colunas$lasso_tab1 <- setdiff(colunas$lasso_tab1,names(
   colunas_nan1) )

colunas$stepwise <- c("DPO","MFM","BIAS","AR_POS","STRSI","DID
    ","ROC","MOM","ARATIO","PPOH","VMOM","EMV","PVOH","DMI","RVI
```

```
    ","SS1","MQO_PRED","AB_DOWN","CHAND_SHORT","APO","VARR","FR1
    ","ATRP","NATR","ULTOSC","DEMA","STOCH_D","DONCHIAN","KST","
    CMO","RMF","OSCP","CHAND_LONG","PSK","MIDPOINT","AB_UP","ADX
    ","VAMA","BB_UP","TRIMA","MASS","EMA","DS1","PVOI","KAMA","
    CCI","PERC_B","OBV","TSI","HULL")
colunas$stepwise <- setdiff(colunas$stepwise, names(colunas_nan1
    ))

colunas$stepwise_tab1 <- c("WILL_R","LPR","REX","ARATIO","ADO","
    STOCH_D_SLOW","STOCH_K","VARR","BB_UP","TRIMA","MINN","MPP","
    DMI","WMA","VAMA","CCI","RSI","KC_U","MACD")
colunas$stepwise_tab1 <- setdiff(colunas$stepwise_tab1, names(
    colunas_nan1))

colunas$torneio <- c("WWS","MFM","KC_M","FR2","VOLR","LPR","NVI
    ","STOCH_K","STRSI","TRIX","RSI","ULTOSC","NATR","DIU","CLOSE
    ","ATR","ATRP","MPP","OSCP","FR1","VOLAT","NVOI","DISP","
    AB_UP","HULL","AR_OSC","ADL","DID","OBV","SS2","BB_BW","VMOM
    ","SS1","DONCHIAN","PVOI","MQO_PRED","PPOH","TP","KC_L","KC_U
    ","CHOSC","CHOPPINESS","VOLUME","DPO","ADO","CVOL","MAE_UP","
    PVI","HPR","DS1")
colunas$torneio <- setdiff(colunas$torneio, names(colunas_nan1))

colunas$torneio_tab1 <- c("MAE_LOW","DIU","SMA","STOCH_K","
    BB_LOW","WMA","BRATIO","CLOSE","WILL_R","CCI","LPR","VOLR","
    DID","BIAS","PPO","DISP","HPR","OBV","SAR","NVI")
colunas$torneio_tab1 <- setdiff(colunas$torneio_tab1, names(
    colunas_nan1))


resultados <- list()
##### ========= CENARIO 1 ========== ######
indicadores <- list()
for (i in 1:length(colunas)){

  colunas_sel <- which( colnames(x_train) %in% colunas[[i]] )

  for (k in c('sigmoid') ){
```

```
for(j in c(0,0.3) ){

  model <- keras_model_sequential()

  # Defines the network's architecture
  model %>%
    layer_dense(units = 15, activation = k, input_shape =
       length(colunas[[i]]) ) %>%
    layer_dropout(rate = j) %>%
    layer_dense(units = 15, activation = k) %>%
    layer_dropout(rate = j) %>%
    layer_dense(units = 15, activation = k) %>%
    layer_dropout(rate = j) %>%
    layer_dense(units = 15, activation = k) %>%
    layer_dropout(rate = j) %>%
    layer_dense(units = 2, activation = 'sigmoid')

  # Loss function and optimization method
  model %>%
    compile(
      loss = 'categorical_crossentropy',
      optimizer = optimizer_rmsprop(),
      metrics = c('accuracy')
    )

  # Trains the model
  history <- model %>% fit(
    x_train[ , colunas_sel ], y_train,
    epochs = 400, batch_size = 128,
    validation_split = 0.2
  )

  resultados[[ length(resultados) + 1 ]] <- data.frame(
     Variables  = names(colunas[i]),
   Hidden_layers = 3,
   Dropout     = j,
   Activation = k,
   Accuracy_in = history$metrics$val_acc[[400]],
   Accuracy_out   = model %>%
```

```
            evaluate(x_test[ , colunas_sel ], y_test) %>%
            .$acc
        )

        indicadores[[ length(indicadores) + 1 ]] <- data.frame(
            y_teste = y_test,
            preditos = predict_classes(model, x_test[, colunas_sel
                ]), preco_close = preco_puro )

    }
  }
}

tabela_1_base1_br <- do.call( rbind, resultados )
indicadores1 <- indicadores
save(indicadores1, file = "conf_price_3lay.RData")
write.csv(tabela_1_base1_USA, file = "tabela_1_base1_USA.csv",
    row.names = F)
save(tabela_1_base1_br, file = 'tabela_1_base1_USA.RData')
##### ---------------------- Fim Cenario 1: ----------------------
    ######

## Procedure is analogous for networks with 5 and 7 hidden
    layers

result_base1 <- rbind( tabela_1_base1_br,
                       tabela_2_base1_br,
                       tabela_3_base1_br)

write.csv(result_base1, file = "result_base123.csv", row.names =
    F)
save(result_base1, file = 'result_base123.RData')

### classification evaluation metrics

mega_lista <- c(indicadores1,indicadores2,indicadores3)

acc <- c()
precision <- c()
```

```
recall <- c()
f_score <- c()

for (i in 1:length(mega_lista)){
  p1 <- ((mega_lista[[i]]$y_teste.1*2+mega_lista[[i]]$y_teste.2)
     -2)*(-1)
  p2 <- mega_lista[[i]]$preditos
  t1 <- table(p1,p2)
  if (ncol(t1)==2){
    acc <- c(acc, sum(diag(t1))/sum(t1))
    pp1 <- t1[2,2]/(t1[2,2]+t1[1,2])
    precision <- c(precision, pp1)
    pp2 <- t1[2,2]/(t1[2,2]+t1[2,1])
    recall <- c(recall, pp2)
    f_score <- c(f_score, (((1/pp1)+(1/pp2))/2)^(-1))
  }
  else{
    if (unique(p2)==0){
      acc <- c(acc, sum(diag(t1))/sum(t1))
      precision <- c(precision, 0)
      recall <- c(recall, 0)
      f_score <- c(f_score, 0)
    }
    if (unique(p2)==1){
      acc <- c(acc, sum(diag(t1))/sum(t1))
      pp1 <- t1[2,1]/(t1[2,1]+t1[1,1])
      precision <- c(precision, pp1)
      pp2 <- 1
      recall <- c(recall, 1)
      f_score <- c(f_score, (((1/pp1)+(1/pp2))/2)^(-1))
    }
  }
}

result_all_metrics <- cbind(result_base1,acc,precision,recall,
   f_score)

print(result_all_metrics)
```

```
write.csv2(result_all_metrics,"metrics_USA.csv")

### Strategy profitabilities and transaction costs

mega_lista <- c(indicadores1,indicadores2,indicadores3)

transactions <- c()
ganho <- c()
TC_zero <- c()
TC_BH <- c()
BandH_macro <- as.numeric()

for (ind in 1:length(mega_lista)){
  donde <- mega_lista[[ind]]
  donde$preco_close <- base_completa2$CLOSE[(nrow(base_completa2
     )-(nrow(donde)-1)):nrow(base_completa2)]
  donde$ativo <- base_completa2$ATIVO[(nrow(base_completa2)-(
     nrow(donde)-1)):nrow(base_completa2)]

  transactions_micro <- c()
  ganho_micro <- c()
  TC_zero_micro <- c()
  TC_BH_micro <- c()
  BandH_micro <- c()

    for (nom in 1:length(unique(donde$ativo))){
    teste <- donde[donde$ativo==unique(donde$ativo)[nom],]
    BandH <- as.numeric(teste$preco_close[nrow(teste)]) - as.
       numeric(teste$preco_close[1])

    acao <- c(ifelse(teste$preditos[1]==1, "compra","mantem"))
      for (i in 1: (nrow(teste)-2) ) {
        if ( teste[["preditos"]][i+1] - teste[["preditos"]][i]
           == 1 ) {
          acao[i+1] <- "compra"
        } else if ( teste[["preditos"]][i+1] - teste[["preditos
           "]][i] == -1 ) {
          acao[i+1] <- "vende"
        } else {
```

```
            acao[i+1] <- "mantem"
          }
        }

      # Number of operations
      transactions_micro <- c(transactions_micro, sum(acao %in% c
          ('vende','compra')))

      # Profitability under each model
      mod_op <- data.frame(Acao  = acao,
                           Fator = ifelse(acao == "compra" , -1 ,
                                          ifelse(acao == "vende" ,
                                                 1 , 0)),
                           Preco = teste[["preco_close"]][1:
                              length(teste[["preco_close"]])-1] )

      ganho_micro <- c(ganho_micro, sum(mod_op[["Preco"]] * mod_op
          [["Fator"]]))

      TC_zero_micro <- c(TC_zero_micro, (sum(mod_op[["Preco"]] *
          mod_op[["Fator"]]))/(sum(acao %in% c('vende','compra'))))
      TC_BH_micro <- c(TC_BH_micro, (sum(mod_op[["Preco"]] *
          mod_op[["Fator"]]) - BandH)/(sum(acao %in% c('vende','
          compra'))))
      BandH_micro <- c(BandH_micro, BandH)
      }

    transactions <- c(transactions,mean(transactions_micro))
    ganho <- c(ganho,mean(ganho_micro))
    TC_zero <- c(TC_zero,mean(TC_zero_micro))
    TC_BH <- c(TC_BH,mean(TC_BH_micro))
    BandH_macro <- mean(BandH_micro)
}

result_strategies <- cbind(result_base1[,1:3],ganho,transactions
    ,TC_zero,TC_BH)

result_strategies$TC_zero[which(is.nan(result_strategies$TC_zero
    ))] <- 0
```

```
result_strategies$TC_BH[which(is.nan(result_strategies$TC_BH))]
   <- 0
result_strategies$TC_BH[which(is.infinite(
   result_strategies$TC_BH))] <- 0

print(result_strategies)

write.csv2(result_strategies,"strategies_USA.csv")

#### Procedure is analogous for all other markets
```

Bibliography

ABU-MOSTAFA, Y. S.; ATIYA, A. F. Introduction to financial forecasting. *Applied Intelligence*, Springer, v. 6, n. 3, p. 205–213, 1996. Cited in page 62.

AGUDOV, N. V.; DUBKOV, A. A.; SPAGNOLO, B. Escape from a metastable state with fluctuating barrier. *Physica A: Statistical Mechanics and its Applications*, Elsevier, v. 325, n. 1-2, p. 144–151, 2003. Cited in page 88.

AKEMANN, G.; BAIK, J.; FRANCESCO, P. D. *The Oxford handbook of random matrix theory*. [S.l.]: Oxford University Press, 2011. Cited in page 92.

ALAOUI, M. E. Random matrix theory and portfolio optimization in moroccan stock exchange. *Physica A: Statistical Mechanics and its Applications*, Elsevier, v. 433, p. 92–99, 2015. Cited in page 91.

ALBUQUERQUE, P. H. M. et al. Na era das máquinas, o emprego é de quem?: Estimação da probabilidade de automação de ocupações no brasil. *Texto para Discussão – Instituto de Pesquisa Econômica Aplicada*, v. 2457, 2019. Cited in page 37.

ALHASHEL, B. S.; ALMUDHAF, F. W.; HANSZ, J. A. Can technical analysis generate superior returns in securitized property markets? evidence from east asia markets. *Pacific-Basin Finance Journal*, Elsevier, v. 47, p. 92–108, 2018. Cited 4 times in pages 123, 125, 126, and 127.

ALMAHDI, S.; YANG, S. Y. An adaptive portfolio trading system: A risk-return portfolio optimization using recurrent reinforcement learning with expected maximum drawdown. *Expert Systems with Applications*, Elsevier, v. 87, p. 267–279, 2017. Cited in page 88.

ALOUD, M. et al. Stylized facts of trading activity in the high frequency fx market: An empirical study. *Journal of Finance and Investment Analysis*, Citeseer, v. 2, n. 4, p. 145–183, 2013. Cited in page 50.

ALVES, R. R. *Seleção por torneios nas estimativas de associação entre marcadores SNP's e fenótipos*. 2014. PhD Thesis, Federal University of Lavras. Cited in page 136.

ANDERSEN, T. G.; BOLLERSLEV, T. Deutsche mark–dollar volatility: intraday activity patterns, macroeconomic announcements, and longer run dependencies. *the Journal of Finance*, Wiley Online Library, v. 53, n. 1, p. 219–265, 1998. Cited 2 times in pages 50 and 63.

ANDERSEN, T. G. et al. *Realized Volatility and Correlation*. 1999. Cited in page 53.

ANDRYCHOWICZ, M. et al. Fair two-party computations via bitcoin deposits. In: SPRINGER. *International Conference on Financial Cryptography and Data Security*. [S.l.], 2014. p. 105–121. Cited in page 56.

ANG, K. K.; QUEK, C. Stock trading using rspop: A novel rough set-based neuro-fuzzy approach. *IEEE Trans. Neural Networks*, v. 17, n. 5, p. 1301–1315, 2006. Cited in page 125.

ARIELY, D.; BERNS, G. S. Neuromarketing: the hope and hype of neuroimaging in business. *Nature reviews neuroscience*, Nature Publishing Group, v. 11, n. 4, p. 284, 2010. Cited in page 34.

ARMANO, G.; MARCHESI, M.; MURRU, A. A hybrid genetic-neural architecture for stock indexes forecasting. *Information Sciences*, Elsevier, v. 170, n. 1, p. 3–33, 2005. Cited 2 times in pages 125 and 126.

ARNTZ, M.; GREGORY, T.; ZIERAHN, U. The risk of automation for jobs in oecd countries. OECD iLibrary, 2016. Cited in page 31.

ATHEY, S. Machine learning and causal inference for policy evaluation. In: ACM. *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. [S.l.], 2015. p. 5–6. Cited in page 33.

ATHEY, S. et al. *Ensemble Methods for Causal Effects in Panel Data Settings*. [S.l.], 2019. Cited in page 33.

AUDRINO, F.; HUANG, C.; OKHRIN, O. Flexible har model for realized volatility. University of St. Gallen, 2016. Cited in page 51.

AUDRINO, F.; KNAUS, S. D. Lassoing the har model: A model selection perspective on realized volatility dynamics. *Econometric Reviews*, Taylor & Francis, v. 35, n. 8-10, p. 1485–1521, 2016. Cited in page 51.

AWARTANI, B. M.; CORRADI, V. Predicting the volatility of the s&p-500 stock index via garch models: the role of asymmetries. *International Journal of Forecasting*, Elsevier, v. 21, n. 1, p. 167–183, 2005. Cited in page 65.

BAEK, C.; ELBECK, M. Bitcoins as an investment or speculative vehicle? a first look. *Applied Economics Letters*, Taylor & Francis, v. 22, n. 1, p. 30–34, 2015. Cited in page 79.

BAI, J.; GHYSELS, E.; WRIGHT, J. H. State space models and midas regressions. *Econometric Reviews*, Taylor & Francis, v. 32, n. 7, p. 779–813, 2013. Cited in page 52.

BAI, J.; SHI, S. *Estimating high dimensional covariance matrices and its applications*. 2011. 199–215 p. Cited in page 97.

BALTA, S. et al. "DinarDirham" Whitepaper. Cited in page 59.

BAN, G.-Y.; KAROUI, N. E.; LIM, A. E. Machine learning and portfolio optimization. *Management Science*, INFORMS, 2016. Cited 2 times in pages 49 and 90.

BARFUSS, W. et al. Parsimonious modeling with information filtering networks. *Physical Review E*, APS, v. 94, n. 6, p. 062306, 2016. Cited 2 times in pages 48 and 90.

BARUNÍK, J.; KŘEHLÍK, T. Combining high frequency data with non-linear models for forecasting energy market volatility. *Expert Systems With Applications*, Elsevier, v. 55, p. 222–242, 2016. Cited 2 times in pages 63 and 64.

BELKE, A.; SETZER, R. Contagion, herding and exchange-rate instability—a survey. *Intereconomics*, Springer, v. 39, n. 4, p. 222–228, 2004. Cited in page 61.

BENGTSSON, C.; HOLST, J. *On portfolio selection: Improved covariance matrix estimation for Swedish asset returns*. [S.l.]: Univ., 2002. Cited in page 96.

BEZERRA, P. C. S.; ALBUQUERQUE, P. H. M. Volatility forecasting via svr–garch with mixture of gaussian kernels. *Computational Management Science*, Springer, v. 14, n. 2, p. 179–196, 2017. Cited 2 times in pages 64 and 80.

BILLIO, M. et al. Econometric measures of connectedness and systemic risk in the finance and insurance sectors. *Journal of financial economics*, Elsevier, v. 104, n. 3, p. 535–559, 2012. Cited in page 96.

BLACK, F.; LITTERMAN, R. Global portfolio optimization. *Financial analysts journal*, CFA Institute, v. 48, n. 5, p. 28–43, 1992. Cited 2 times in pages 86 and 94.

BLACK, F.; SCHOLES, M. The pricing of options and corporate liabilities. *Journal of political economy*, The University of Chicago Press, v. 81, n. 3, p. 637–654, 1973. Cited in page 87.

BLOCKCHAIN. *Bitcoin - Confirmed Transactions Per Day*. 2017. https://blockchain.info/charts/n-transactions. [Online; accessed 16-March-2017]. Cited in page 57.

BLOCKCHAIN. *Bitcoin - Market Capitalization*. 2017. https://blockchain.info/en/charts/market-cap. [Online; accessed 16-March-2017]. Cited in page 58.

BOLLERSLEV, T. Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, Elsevier, v. 31, n. 3, p. 307–327, 1986. Cited 2 times in pages 63 and 87.

BOLLERSLEV, T.; PATTON, A. J.; QUAEDVLIEG, R. Exploiting the errors: A simple approach for improved volatility forecasting. *Journal of Econometrics*, Elsevier, v. 192, n. 1, p. 1–18, 2016. Cited in page 52.

BONANNO, G.; VALENTI, D.; SPAGNOLO, B. Role of noise in a market model with stochastic volatility. *The European Physical Journal B-Condensed Matter and Complex Systems*, Springer, v. 53, n. 3, p. 405–409, 2006. Cited in page 87.

BOSCH, M.; PAGÉS, C.; RIPANI, L. *El futuro del trabajo en América Latina y el Caribe:¿ Una gran oportunidad para la región*. [S.l.]: Inter-American Development Bank, 2018. Cited 2 times in pages 31 and 37.

BOUCHAUD, J.-P.; POTTERS, M. Financial applications of random matrix theory: a short review. *arXiv preprint arXiv:0910.1205*, 2009. Cited in page 92.

BOURI, E.; AZZI, G.; DYHRBERG, A. H. *On the return-volatility relationship in the Bitcoin market around the price crash of 2013*. [S.l.], 2016. Cited in page 59.

BOURI, E. et al. *Does Bitcoin Hedge Global Uncertainty? Evidence from Wavelet-Based Quantile-in-Quantile Regressions*. [S.l.], 2016. Cited in page 58.

BOYD, D.; CRAWFORD, K. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, Taylor & Francis, v. 15, n. 5, p. 662–679, 2012. Cited 2 times in pages 30 and 34.

BROCK, W. A. Causality, chaos, explanation and prediction in economics and finance. In: *Beyond Belief*. [S.l.]: CRC Press, 2018. p. 230–279. Cited in page 42.

BROCK, W. A. Nonlinearity and complex dynamics in economics and finance. In: *The economy as an evolving complex system*. [S.l.]: CRC Press, 2018. p. 77–97. Cited in page 42.

BROWN, S.; GOETZMANN, W.; ROSS, S. Survivorship bias in performance studies. *Review of Financial Studies*, v. 5, n. 4, p. 553–580, 1992. ISSN 14657368. Disponível em: <http://rfs.oupjournals.org/cgi/doi/10.1093/rfs/5.4.553>. Cited in page 113.

BRYNJOLFSSON, E.; MCAFEE, A. The big data boom is the innovation story of our time. *The Atlantic*, v. 21, 2011. Cited in page 30.

BRYNJOLFSSON, E.; MCAFEE, A. *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. [S.l.]: WW Norton & Company, 2014. Cited in page 31.

BUONOCORE, R. et al. Two different flavours of complexity in financial data. *The European Physical Journal Special Topics*, Springer, v. 225, n. 17-18, p. 3105–3113, 2016. Cited 2 times in pages 42 and 86.

BURDA, Z. et al. Signal and noise in correlation matrix. *Physica A: Statistical Mechanics and its Applications*, Elsevier, v. 343, p. 295–310, 2004. Cited in page 93.

BURNS, K.; MOOSA, I. A. Enhancing the forecasting power of exchange rate models by introducing nonlinearity: Does it work? *Economic Modelling*, Elsevier BV, v. 50, p. 27–39, Nov 2015. ISSN 0264-9993. Cited in page 41.

CAMARGO, S.; QUEIROS, S. M. D.; ANTENEODO, C. Bridging stylized facts in finance and data non-stationarities. *The European Physical Journal B*, Springer, v. 86, n. 4, p. 159, 2013. Cited in page 50.

CAO, L.-J.; TAY, F. E. H. Support vector machine with adaptive parameters in financial time series forecasting. *IEEE Transactions on neural networks*, IEEE, v. 14, n. 6, p. 1506–1518, 2003. Cited in page 62.

CARAYANNIS, E. G.; SINDAKIS, S.; WALTER, C. Business model innovation as lever of organizational sustainability. *The Journal of Technology Transfer*, Springer, v. 40, n. 1, p. 85–104, 2015. Cited in page 37.

CATOR, E. A.; LOPUHAÄ, H. P. Central limit theorem and influence function for the mcd estimators at general multivariate distributions. *Bernoulli*, JSTOR, p. 520–551, 2012. Cited in page 95.

Cavendish Astrophysics. *Technical Indicators and Overlays*. 2011. https://www.mrao.cam.ac.uk/~mph/Technical_Analysis.pdf. Cited 2 times in pages 128 and 130.

ÇELIK, S.; ERGIN, H. Volatility forecasting using high frequency data: Evidence from stock markets. *Economic modelling*, Elsevier, v. 36, p. 176–190, 2014. Cited in page 63.

CHANDRASHEKAR, G.; SAHIN, F. A survey on feature selection methods. *Computers & Electrical Engineering*, Elsevier, v. 40, n. 1, p. 16–28, 2014. Cited 2 times in pages 134 and 135.

CHANG, C.-C.; LIN, C.-J. Training ν -support vector regression: Theory and algorithms. *Neural Computation*, v. 14, n. 8, p. 1959–1977, 2002. Cited 2 times in pages 66 and 80.

CHANG, P.-C.; FAN, C.-Y. A hybrid system integrating a wavelet and tsk fuzzy rules for stock price forecasting. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, IEEE, v. 38, n. 6, p. 802–815, 2008. Cited 2 times in pages 125 and 126.

CHANG, P.-C. et al. A neural network with a case based dynamic window for stock trading prediction. *Expert Systems with Applications*, Elsevier, v. 36, n. 3, p. 6889–6898, 2009. Cited 3 times in pages 125, 126, and 127.

CHANG, P.-C. et al. A novel model by evolving partially connected neural network for stock price trend forecasting. *Expert Systems with Applications*, Elsevier, v. 39, n. 1, p. 611–620, 2012. Cited 2 times in pages 125 and 126.

CHAO, J.; SHEN, F.; ZHAO, J. Forecasting exchange rate with deep belief networks. In: IEEE. *International Joint Conference on Neural Networks (IJCNN)*. [S.l.], 2011. p. 1259–1266. Cited in page 41.

CHAUM, D. L. Blind signatures for untraceable payments. 1983. Online; access 07-November-2017. Disponível em: <http://ecash.com/>. Cited in page 56.

CHAVARNAKUL, T.; ENKE, D. A hybrid stock trading system for intelligent technical analysis-based equivolume charting. *Neurocomputing*, Elsevier, v. 72, n. 16-18, p. 3517–3528, 2009. Cited in page 126.

CHEN, C.; ZHOU, Y.-s. Robust multiobjective portfolio with higher moments. *Expert Systems with Applications*, Elsevier, v. 100, p. 165–181, 2018. Cited in page 87.

CHEN, C.-C. et al. Applying market profile theory to forecast taiwan index futures market. *Expert Systems with Applications*, Elsevier, v. 41, n. 10, p. 4617–4624, 2014. Cited in page 125.

CHEN, H.; CHIANG, R. H.; STOREY, V. C. Business intelligence and analytics: From big data to big impact. *MIS quarterly*, v. 36, n. 4, 2012. Cited in page 30.

CHEN, H. et al. A double-layer neural network framework for high-frequency forecasting. *ACM Transactions on Management Information Systems (TMIS)*, ACM, v. 7, n. 4, p. 11, 2017. Cited in page 125.

CHEN, J.-S. et al. Constructing investment strategy portfolios by combination genetic algorithms. *Expert Systems with Applications*, Elsevier, v. 36, n. 2, p. 3824–3828, 2009. Cited in page 88.

CHEN, S.; HÄRDLE, W. K.; JEONG, K. Forecasting volatility with support vector machine-based garch model. *Journal of Forecasting*, Wiley Online Library, v. 29, n. 4, p. 406–433, 2010. Cited 2 times in pages 64 and 66.

CHEN, S.; JEONG, K.; HÄRDLE, W. K. Support vector regression based garch model with application to forecasting volatility of financial returns. 2008. Cited 2 times in pages 62 and 63.

CHEN, Y.-C.; HSIEH, T.-C. Big data for digital government: opportunities, challenges, and strategies. *International journal of public administration in the digital age (IJPADA)*, IGI Global, v. 1, n. 1, p. 1–14, 2014. Cited in page 33.

CHEN, Y.-S.; CHENG, C.-H.; TSAI, W.-L. Modeling fitting-function-based fuzzy time series patterns for evolving stock index forecasting. *Applied intelligence*, Springer, v. 41, n. 2, p. 327–347, 2014. Cited 2 times in pages 125 and 126.

CHEN, Z.; CHEN, J. Tournament screening cum ebic for feature selection with high-dimensional feature spaces. *Science in China Series A: Mathematics*, Springer, v. 52, n. 6, p. 1327–1341, 2009. Cited in page 135.

CHIANG, W.-C. et al. An adaptive stock index trading decision support system. *Expert Systems with Applications*, Elsevier, v. 59, p. 195–207, 2016. Cited 2 times in pages 125 and 127.

CHICHEPORTICHE, R.; BOUCHAUD, J.-P. A nested factor model for non-linear dependencies in stock returns. *Quantitative Finance*, Taylor & Francis, v. 15, n. 11, p. 1789–1804, 2015. Cited in page 89.

COCHRANE, J. H. Presidential address: Discount rates. *The Journal of finance*, Wiley Online Library, v. 66, n. 4, p. 1047–1108, 2011. Cited in page 117.

COHEN, C. Beenz.com; the web's currency. 1998. Online; access 07-November-2017. Disponível em: <https://beenz.com/>. Cited in page 56.

COINDESK. *Bitcoin - Daily Number of Transactions*. 2017. http://www.coindesk.com/data/bitcoin-daily-transactions/. [Online; accessed 16-March-2017]. Cited in page 57.

COINDESK. *Bitcoin - Market Capitalization*. 2017. http://www.coindesk.com/data/bitcoin-market-capitalization/. [Online; accessed 16-March-2017]. Cited in page 58.

COMMUNITY, B. bitcoinwiki. 2017. Online; access 11-November-2017. Disponível em: <https://en.bitcoin.it/wiki/Confirmation>. Cited in page 56.

CONLON, T.; RUSKIN, H. J.; CRANE, M. Random matrix theory and fund of funds portfolio optimisation. *Physica A: Statistical Mechanics and its Applications*, v. 382, n. 2, p. 565–576, 2007. ISSN 03784371. Cited 3 times in pages 92, 97, and 98.

CONRAD, C.; LAMLA, M. J. The high-frequency response of the EUR-USD exchange rate to ECB communication. *Journal of Money, Credit and Banking*, Wiley Online Library, v. 42, n. 7, p. 1391–1417, 2010. Cited in page 41.

CONSIGLIO, A.; CAROLLO, A.; ZENIOS, S. A. A parsimonious model for generating arbitrage-free scenario trees. *Quantitative Finance*, Taylor & Francis, v. 16, n. 2, p. 201–212, 2016. Cited in page 93.

CONT, R. Empirical properties of asset returns: stylized facts and statistical issues. Taylor & Francis, 2001. Cited 2 times in pages 38 and 65.

CORSI, F. A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, Oxford University Press, v. 7, n. 2, p. 174–196, 2009. Cited in page 51.

CORTES, C.; VAPNIK, V. N. Support-vector networks. *Machine Learning*, v. 20, p. 273–297, 1995. Cited in page 39.

COSTA, T. R. C. C. da et al. Trading system based on the use of technical analysis: A computational experiment. *Journal of Behavioral and Experimental Finance*, Elsevier, v. 6, p. 42–55, 2015. Cited in page 122.

CREAMER, G. Model calibration and automated trading agent for euro futures. *Quantitative Finance*, Taylor & Francis, v. 12, n. 4, p. 531–545, 2012. Cited 4 times in pages 118, 125, 126, and 127.

CROUX, C.; HAESBROECK, G. Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. *Journal of Multivariate Analysis*, Academic press inc, v. 71, n. 2, p. 161–190, 1999. Cited in page 95.

DALY, J.; CRANE, M.; RUSKIN, H. J. Random matrix theory filters in portfolio optimisation: a stability and risk assessment. *Physica A: Statistical Mechanics and its Applications*, Elsevier, v. 387, n. 16-17, p. 4248–4260, 2008. Cited in page 114.

DAMODARAN, A. *Investment valuation: Tools and techniques for determining the value of any asset*. [S.l.]: John Wiley & Sons, 2012. Cited in page 85.

DARLINGTON, J. The future of bitcoin: Mapping the global adoption of world's largest cryptocurrency through benefit analysis. 2014. Cited in page 56.

DAVENPORT, T. *Big data at work: dispelling the myths, uncovering the opportunities*. [S.l.]: Harvard Business Review Press, 2014. Cited in page 36.

DEBOECK, G. *Trading on the edge: neural, genetic, and fuzzy systems for chaotic financial markets*. [S.l.]: John Wiley & Sons, 1994. Cited in page 62.

DESOUZA, K. C.; JACOB, B. Big data in the public sector: Lessons for practitioners and scholars. *Administration & Society*, SAGE Publications Sage CA: Los Angeles, CA, v. 49, n. 7, p. 1043–1064, 2017. Cited in page 34.

DIEBOLD, F. X.; MARIANO, R. S. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, Informa UK Limited, v. 13, n. 3, p. 253–263, Jul 1995. ISSN 1537-2707. Cited 4 times in pages 54, 63, 70, and 81.

DING, Z.; GRANGER, C. W.; ENGLE, R. F. A long memory property of stock market returns and a new model. *Journal of empirical finance*, Elsevier, v. 1, n. 1, p. 83–106, 1993. Cited in page 80.

DOWD, K. New private monies: A bit-part player? 2014. Cited 2 times in pages 60 and 79.

DRIESSEN, J.; MELENBERG, B.; NIJMAN, T. Common factors in international bond returns. *Journal of International Money and Finance*, Elsevier, v. 22, n. 5, p. 629–656, 2003. Cited in page 96.

DRUCKER, H. et al. Support vector regression machines. In: *Advances in neural information processing systems*. [S.l.]: Morgan Kaufmann Publishers, 1997. p. 155–161. Cited 2 times in pages 66 and 67.

DUBKOV, A. A.; AGUDOV, N. V.; SPAGNOLO, B. Noise-enhanced stability in fluctuating metastable states. *Physical Review E*, APS, v. 69, n. 6, p. 061103, 2004. Cited in page 88.

DYHRBERG, A. H. Bitcoin, gold and the dollar–a garch volatility analysis. *Finance Research Letters*, Elsevier, v. 16, p. 85–92, 2016. Cited 4 times in pages 58, 59, 60, and 80.

DYHRBERG, A. H. Hedging capabilities of bitcoin. is it the virtual gold? *Finance Research Letters*, Elsevier, v. 16, p. 139–144, 2016. Cited in page 58.

EASLEY, D.; PRADO, M. Lopez de; O'HARA, M. The volume clock: Insights into the high frequency paradigm. 2012. Cited 4 times in pages 49, 50, 62, and 63.

EDELMAN, A. Eigenvalues and condition numbers of random matrices. *SIAM Journal on Matrix Analysis and Applications*, SIAM, v. 9, n. 4, p. 543–560, 1988. Cited in page 91.

EFRON, B.; HASTIE, T. *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. [S.l.]: Cambridge University Press, 2016. Cited in page 32.

ENGLE, R. F. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica: Journal of the Econometric Society*, JSTOR, p. 987–1007, 1982. Cited in page 63.

EREVELLES, S.; FUKAWA, N.; SWAYNE, L. Big data consumer analytics and the transformation of marketing. *Journal of Business Research*, Elsevier, v. 69, n. 2, p. 897–904, 2016. Cited in page 34.

ETEROVIC, N. *A Random Matrix Approach to Portfolio Management and Financial Networks*. 2016. PhD Thesis, University of Essex. Cited in page 92.

ETEROVIC, N. A.; ETEROVIC, D. S. Separating the wheat from the chaff: Understanding portfolio returns in an emerging market. *Emerging Markets Review*, Elsevier, v. 16, p. 145–169, 2013. Cited in page 92.

EVANS, C.; PAPPAS, K.; XHAFA, F. Utilizing artificial neural networks and genetic algorithms to build an algo-trading model for intra-day foreign exchange speculation. *Mathematical and Computer Modelling*, Elsevier, v. 58, n. 5, p. 1249–1266, 2013. Cited in page 62.

EVANS, D. S. Economic aspects of bitcoin and other decentralized public-ledger currency platforms. 2014. Cited in page 57.

FAMA, E. F. Efficient capital markets: A review of theory and empirical work. *The journal of finance*, Wiley Online Library, v. 25, n. 2, p. 383–417, 1970. Cited in page 39.

FAMA, E. F.; FRENCH, K. R. The cross-section of expected stock returns. *the Journal of Finance*, Wiley Online Library, v. 47, n. 2, p. 427–465, 1992. Cited 2 times in pages 48 and 118.

FAMA, E. F.; FRENCH, K. R. Multifactor explanations of asset pricing anomalies. *The journal of finance*, Wiley Online Library, v. 51, n. 1, p. 55–84, 1996. Cited in page 118.

FAMA, E. F.; FRENCH, K. R. A five-factor asset pricing model. *Journal of Financial Economics*, v. 116, n. 1, p. 1 – 22, 2015. ISSN 0304-405X. Disponível em: <http://www-.sciencedirect.com/science/article/pii/S0304405X14002323>. Cited 2 times in pages 48 and 118.

FAN, J.; LI, R. Statistical challenges with high dimensionality: Feature selection in knowledge discovery. *arXiv preprint math/0602133*, 2006. Cited in page 116.

FAN, J.; LV, J. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, NIH Public Access, v. 20, n. 1, p. 101, 2010. Cited in page 116.

FARAGO, A.; TÉDONGAP, R. Downside risks and the cross-section of asset returns. *Journal of Financial Economics*, Elsevier, 2018. Cited in page 101.

FENG, G.; GIGLIO, S.; XIU, D. Taming the factor zoo. 2017. Cited in page 118.

FENG, G.; POLSON, N. G.; XU, J. Deep factor alpha. *arXiv preprint arXiv:1805.01104*, 2018. Cited in page 48.

FENGHUA, W. et al. Stock price prediction based on ssa and svm. *Procedia Computer Science*, Elsevier, v. 31, p. 625–631, 2014. Cited in page 120.

FERGUSON, N. *The ascent of money: A financial history of the world*. [S.l.]: Penguin, 2008. Cited in page 55.

FIASCONARO, A.; SPAGNOLO, B.; BOCCALETTI, S. Signatures of noise-enhanced stability in metastable states. *Physical Review E*, APS, v. 72, n. 6, p. 061110, 2005. Cited in page 88.

Fidelity Investments. *Technical Indicator Guide*. 2019. https://www.fidelity.com/learning-center/trading-investing/technical-analysis/technical-indicator-guide/overview. Cited 3 times in pages 128, 129, and 130.

FOSTER, I. et al. *Big data and social science: A practical guide to methods and tools*. [S.l.]: Chapman and Hall/CRC, 2016. Cited in page 30.

FRAHM, G.; JAEKEL, U. Random matrix theory and robust covariance matrix estimation for financial data. *arXiv preprint*, p. 1–22, 2005. Disponível em: <http://arxiv.org/abs/physics/0503007>. Cited in page 98.

FREY, C. B. et al. Technology at work v2. 0: The future is not what it used to be. *CityGroup and University of Oxford*, 2016. Cited in page 31.

FREY, C. B.; OSBORNE, M. A. The future of employment: how susceptible are jobs to computerisation? *Technological forecasting and social change*, Elsevier, v. 114, p. 254–280, 2017. Cited in page 31.

GALLOPPO, G. et al. A comparison of pre and post modern portfolio theory using resampling. *Global Journal of Business Research*, The Institute for Business and Finance Research, v. 4, n. 1, p. 1–16, 2010. Cited in page 86.

GANDAL, N.; HALABURDA, H. Competition in the cryptocurrency market. 2014. Cited in page 57.

GANDOMI, A.; HAIDER, M. Beyond the hype: Big data concepts, methods, and analytics. *International journal of information management*, Elsevier, v. 35, n. 2, p. 137–144, 2015. Cited in page 30.

GAVRISHCHAKA, V. V.; BANERJEE, S. Support vector machine as an efficient framework for stock market volatility forecasting. *Computational Management Science*, Springer, v. 3, n. 2, p. 147–160, 2006. Cited 2 times in pages 62 and 63.

GAVRISHCHAKA, V. V.; GANGULI, S. B. Volatility forecasting from multiscale and high-dimensional market data. *Neurocomputing*, Elsevier, v. 55, n. 1, p. 285–305, 2003. Cited 2 times in pages 62 and 63.

GELMAN, A. et al. Induction and deduction in bayesian data analysis. *Rationality, Markets and Morals*, Frankfurt School Verlag, Frankfurt School of Finance & Management, v. 2, n. 67-78, p. 1999, 2011. Cited in page 32.

GENTON, M. G. Classes of Kernels for Machine Learning: A Statistics Perspective. *Journal of Machine Learning Research*, v. 2, p. 299–312, 2001. ISSN 15324435. Cited in page 113.

GEORGE, G.; HAAS, M. R.; PENTLAND, A. *Big data and management.* [S.l.]: Academy of Management Briarcliff Manor, NY, 2014. Cited in page 32.

GERLEIN, E. A. et al. Evaluating machine learning classification for financial trading: An empirical approach. *Expert Systems with Applications*, Elsevier BV, v. 54, p. 193–207, Jul 2016. ISSN 0957-4174. Cited in page 39.

GEYER, A.; HANKE, M.; WEISSENSTEINER, A. No-arbitrage rom simulation. *Journal of Economic Dynamics and Control*, Elsevier, v. 45, p. 66–79, 2014. Cited in page 93.

GHYSELS, E.; SANTA-CLARA, P.; VALKANOV, R. The midas touch: Mixed data sampling regression models. 2004. Cited in page 52.

GLASS, R.; CALLAHAN, S. *The Big Data-driven business: How to use big data to win customers, beat competitors, and boost profits.* [S.l.]: John Wiley & Sons, 2014. Cited in page 34.

GLOSTEN, L. R.; JAGANNATHAN, R.; RUNKLE, D. E. On the relation between the expected value and the volatility of the nominal excess return on stocks. *The journal of finance*, Wiley Online Library, v. 48, n. 5, p. 1779–1801, 1993. Cited in page 65.

GNANADESIKAN, R.; KETTENRING, J. R. Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, JSTOR, p. 81–124, 1972. Cited in page 95.

GOBBLE, M. M. Big data: The next big thing in innovation. *Research-technology management*, Taylor & Francis, v. 56, n. 1, p. 64–67, 2013. Cited in page 37.

GOLOSNOY, V.; GRIBISCH, B.; LIESENFELD, R. The conditional autoregressive wishart model for multivariate stock market volatility. *Journal of Econometrics*, Elsevier, v. 167, n. 1, p. 211–223, 2012. Cited in page 52.

GOMBER, P.; HAFERKORN, M. High frequency trading. In: *Encyclopedia of Information Science and Technology, Third Edition.* [S.l.]: IGI Global, 2015. p. 1–9. Cited in page 50.

GOODFELLOW, I. et al. *Deep learning*. [S.l.]: MIT press Cambridge, 2016. Cited in page 132.

G&SR. E-gold; better money; since 1996. 1998. Online; access 07-November-2017. Disponível em: <https://web.archive.org/web/19981202142942/http://www.e-gold-.com:80/>. Cited in page 56.

G&SR. What is e-gold? 2006. Online; access 07-November-2017. Disponível em: <https:-//web.archive.org/web/20061109064453/http://e-gold.com:80/unsecure/qanda.html>. Cited in page 56.

G&SR. What is e-gold? 2006. Online; access 07-November-2017. Disponível em: <https://web.archive.org/web/20061109161419/http://www.e-gold.com/stats.html>. Cited in page 56.

GU, S.; KELLY, B. T.; XIU, D. Empirical asset pricing via machine learning. 2018. Cited 3 times in pages 47, 121, and 133.

GUHR, T.; KÄLBER, B. A new method to estimate the noise in financial correlation matrices. *Journal of Physics A: Mathematical and General*, IOP Publishing, v. 36, n. 12, p. 3009, 2003. Cited in page 114.

GUNASEKARAN, A. et al. Big data and predictive analytics for supply chain and organizational performance. *Journal of Business Research*, Elsevier, v. 70, p. 308–317, 2017. Cited in page 35.

GÜNTHER, W. A. et al. Debating big data: A literature review on realizing value from big data. *The Journal of Strategic Information Systems*, Elsevier, v. 26, n. 3, p. 191–209, 2017. Cited in page 32.

GUPTA, P.; MEHLAWAT, M. K.; MITTAL, G. Asset portfolio optimization using support vector machines and real-coded genetic algorithm. *Journal of Global Optimization*, Springer, v. 53, n. 2, p. 297–315, 2012. Cited 2 times in pages 43 and 88.

GURESEN, E.; KAYAKUTLU, G.; DAIM, T. U. Using artificial neural network models in stock market index prediction. *Expert Systems with Applications*, Elsevier, v. 38, n. 8, p. 10389–10397, 2011. Cited in page 119.

GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. *Journal of machine learning research*, v. 3, n. Mar, p. 1157–1182, 2003. Cited in page 120.

HAN, Q. et al. Modeling nonlinearity in multi-dimensional dependent data. In: IEEE. *Signal and Information Processing (GlobalSIP), 2017 IEEE Global Conference on*. [S.l.], 2017. p. 206–210. Cited in page 42.

HANSEN, P. R.; LUNDE, A. A forecast comparison of volatility models: does anything beat a garch (1, 1)? *Journal of applied econometrics*, Wiley Online Library, v. 20, n. 7, p. 873–889, 2005. Cited 3 times in pages 63, 64, and 65.

HANSEN, P. R.; LUNDE, A.; NASON, J. M. The model confidence set. *Econometrica*, Wiley Online Library, v. 79, n. 2, p. 453–497, 2011. Cited 8 times in pages 54, 70, 71, 72, 73, 74, 78, and 81.

HARVEY, C. R.; LIU, Y.; ZHU, H. ... and the cross-section of expected returns. *The Review of Financial Studies*, Oxford University Press, v. 29, n. 1, p. 5–68, 2016. Cited in page 117.

HEATON, J.; POLSON, N.; WITTE, J. H. Deep learning for finance: deep portfolios. *Applied Stochastic Models in Business and Industry*, Wiley Online Library, v. 33, n. 1, p. 3–12, 2017. Cited 4 times in pages 43, 88, 132, and 133.

HEATON, J.; POLSON, N. G.; WITTE, J. H. Deep learning in finance. *arXiv preprint arXiv:1602.06561*, 2016. Cited in page 132.

HECKLAU, F. et al. Holistic approach for human resource management in industry 4.0. *Procedia Cirp*, Elsevier, v. 54, p. 1–6, 2016. Cited in page 37.

HENRIQUE, B. M.; SOBREIRO, V. A.; KIMURA, H. Stock price prediction using support vector regression on daily and up to the minute prices. *The Journal of Finance and Data Science*, Elsevier, v. 4, n. 3, p. 183–201, 2018. Cited 3 times in pages 122, 125, and 126.

HENRIQUE, B. M.; SOBREIRO, V. A.; KIMURA, H. Literature review: Machine learning techniques applied to financial market prediction. *Expert Systems with Applications*, Elsevier, 2019. Cited 3 times in pages 122, 133, and 138.

HENRIQUE, P. A. et al. Portfolio selection with support vector regression. In: *R Finance Chicago*. [S.l.: s.n.], 2016. Cited in page 44.

HESTON, S. L. A closed-form solution for options with stochastic volatility with applications to bond and currency options. *The review of financial studies*, Oxford University Press, v. 6, n. 2, p. 327–343, 1993. Cited in page 87.

HOEFFDING, W. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, Taylor & Francis Group, v. 58, n. 301, p. 13–30, 1963. Cited in page 45.

HSU, M.-W. et al. Bridging the divide in financial market forecasting: machine learners vs. financial economists. *Expert Systems with Applications*, Elsevier, v. 61, p. 215–234, 2016. Cited 7 times in pages 39, 43, 64, 79, 84, 89, and 131.

HUANG, C.-F. A hybrid stock selection model using genetic algorithms and support vector regression. *Applied Soft Computing*, Elsevier, v. 12, n. 2, p. 807–818, 2012. Cited in page 86.

HUANG, C.-L.; TSAI, C.-Y. A hybrid sofm-svr with a filter-based feature selection for stock market forecasting. *Expert Systems with Applications*, Elsevier, v. 36, n. 2, p. 1529–1539, 2009. Cited 2 times in pages 125 and 126.

HUBERT, M.; DRIESSEN, K. V. Fast and robust discriminant analysis. *Computational Statistics & Data Analysis*, Elsevier, v. 45, n. 2, p. 301–320, 2004. Cited in page 95.

HUDA, M. et al. Big data emerging technology: insights into innovative environment for online learning resources. *International Journal of Emerging Technologies in Learning (iJET)*, International Association of Online Engineering, v. 13, n. 1, p. 23–36, 2018. Cited in page 37.

HUO, L.; KIM, T.-H.; KIM, Y. Robust estimation of covariance and its application to portfolio optimization. *Finance Research Letters*, Elsevier, v. 9, n. 3, p. 121–134, 2012. Cited in page 94.

HWANG, S.; RUBESAM, A. Searching the factor zoo. 2018. Cited in page 117.

JENSSEN, R. Kernel entropy component analysis. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 32, n. 5, p. 847–860, 2010. Cited in page 97.

JOHN, G. H.; KOHAVI, R.; PFLEGER, K. Irrelevant features and the subset selection problem. In: *Machine Learning Proceedings 1994*. [S.l.]: Elsevier, 1994. p. 121–129. Cited in page 134.

KALMAN, R. E. A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, American Society of Mechanical Engineers, v. 82, n. 1, p. 35–45, 1960. Cited in page 52.

KANAS, A. Nonlinearity in the stock price–dividend relation. *Journal of International Money and Finance*, Elsevier, v. 24, n. 4, p. 583–606, 2005. Cited in page 41.

KARA, Y.; BOYACIOGLU, M. A.; BAYKAN, Ö. K. Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the istanbul stock exchange. *Expert systems with Applications*, Elsevier, v. 38, n. 5, p. 5311–5319, 2011. Cited 2 times in pages 125 and 126.

KARIMI, H. A. *Big Data: techniques and technologies in geoinformatics*. [S.l.]: Crc Press, 2014. Cited in page 35.

KAROLYI, G. A. *Cracking the Emerging Markets Enigma*. [S.l.]: Oxford University Press, 2015. Cited in page 111.

KIM, D.-H.; JEONG, H. Systematic analysis of group identification in stock markets. *Physical Review E*, APS, v. 72, n. 4, p. 046133, 2005. Cited in page 96.

KIM, G.-H.; TRIMI, S.; CHUNG, J.-H. Big-data applications in the government sector. *Communications of the ACM*, ACM, v. 57, n. 3, p. 78–85, 2014. Cited in page 34.

KIM, K.-j. Financial time series forecasting using support vector machines. *Neurocomputing*, Elsevier, v. 55, n. 1-2, p. 307–319, 2003. Cited 2 times in pages 125 and 126.

KIM, K.-j.; HAN, I. Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index. *Expert systems with Applications*, Elsevier, v. 19, n. 2, p. 125–132, 2000. Cited 2 times in pages 125 and 126.

KIM, T. H. A study of digital currency cryptography for business marketing and finance security. *Asia-pacific Journal of Multimedia Services Convergent with Art, Humanities, and Sociology*, v. 6, n. 1, p. 365–376, 2016. Cited in page 59.

KINGMA, D. P.; BA, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. Cited in page 138.

KOZAK, S.; NAGEL, S.; SANTOSH, S. *Shrinking the cross section*. [S.l.], 2017. Cited 3 times in pages 49, 90, and 117.

KRISTOUFEK, L. What are the main drivers of the bitcoin price? evidence from wavelet coherence analysis. *PloS one*, Public Library of Science, v. 10, n. 4, p. e0123923, 2015. Cited 2 times in pages 57 and 58.

KÜHN, R.; NEU, P. Intermittency in an interacting generalization of the geometric brownian motion model. *Journal of Physics A: Mathematical and Theoretical*, IOP Publishing, v. 41, n. 32, p. 324015, 2008. Cited in page 87.

KWON, Y.-K.; MOON, B.-R. A hybrid neurogenetic approach for stock forecasting. *IEEE transactions on neural networks*, IEEE, v. 18, n. 3, p. 851–864, 2007. Cited in page 125.

LALOUX, L. et al. Noise dressing of financial correlation matrices. *Physical review letters*, APS, v. 83, n. 7, p. 1467, 1999. Cited 2 times in pages 91 and 118.

LALOUX, L. et al. Random matrix theory and financial correlations. *International Journal of Theoretical and Applied Finance*, World Scientific, v. 3, n. 03, p. 391–397, 2000. Cited in page 93.

LAURENT, S.; ROMBOUTS, J. V.; VIOLANTE, F. On the forecasting accuracy of multivariate garch models. *Journal of Applied Econometrics*, Wiley Online Library, v. 27, n. 6, p. 934–955, 2012. Cited in page 65.

LAVERTU, S. We all need help:"big data" and the mismeasure of public administration. *Public Administration Review*, Wiley Online Library, v. 76, n. 6, p. 864–872, 2016. Cited in page 33.

LEE, J.; KAO, H.-A.; YANG, S. Service innovation and smart analytics for industry 4.0 and big data environment. *Procedia Cirp*, Elsevier, v. 16, p. 3–8, 2014. Cited in page 37.

LEE, M.-C. Using support vector machine with a hybrid feature selection method to the stock trend prediction. *Expert Systems with Applications*, Elsevier, v. 36, n. 8, p. 10896–10904, 2009. Cited in page 135.

LEVITAN, R. "flooz.com". 1999. Online; posted 23-October-2017. Disponível em: <https://www.theguardian.com/technology/2001/aug/28/newmedia.business>. Cited in page 56.

LI, M.; SUOHAI, F. Forex Prediction Based on SVR Optimized by Artificial Fish Swarm Algorithm. *Fourth Global Congress on Intelligent Systems*, Institute of Electrical & Electronics Engineers (IEEE), p. 47–52, Dec 2013. Cited in page 66.

LI, X.; WANG, C. A. The technology and economic determinants of cryptocurrency exchange rates: The case of bitcoin. *Decision Support Systems*, Elsevier, 2016. Cited 2 times in pages 62 and 63.

LINTNER, J. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *The Review of Economics and Statistics*, JSTOR, p. 13–37, 1965. Cited in page 85.

LIVAN, G.; INOUE, J.-i.; SCALAS, E. On the non-stationarity of financial time series: impact on optimal portfolio selection. *Journal of Statistical Mechanics: Theory and Experiment*, IOP Publishing, v. 2012, n. 07, p. P07025, 2012. Cited 3 times in pages 43, 87, and 91.

LO, S.; WANG, J. C. Bitcoin as money? 2014. Cited in page 60.

MACIEJEWSKI, M. To do more, better, faster and more cheaply: Using big data in public administration. *International Review of Administrative Sciences*, SAGE Publications Sage UK: London, England, v. 83, n. 1_suppl, p. 120–135, 2017. Cited in page 34.

MAMA, H. B. Innovative efficiency and stock returns: Should we care about nonlinearity? *Finance Research Letters*, Elsevier, 2017. Cited in page 42.

MARCELINO, S.; HENRIQUE, P. A.; ALBUQUERQUE, P. H. M. Portfolio selection with support vector machines in low economic perspectives in emerging markets. *Economic Computation & Economic Cybernetics Studies & Research*, v. 49, n. 4, 2015. Cited in page 86.

MARČENKO, V. A.; PASTUR, L. A. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, IOP Publishing, v. 1, n. 4, p. 457, 1967. Cited 2 times in pages 92 and 98.

MARCUCCI, J. et al. Forecasting stock market volatility with regime-switching garch models. *Studies in Nonlinear dynamics and Econometrics*, v. 9, n. 4, p. 1–53, 2005. Cited in page 63.

MARKOWITZ, H. Portfolio selection. *The journal of finance*, Wiley Online Library, v. 7, n. 1, p. 77–91, 1952. Cited 9 times in pages 43, 84, 85, 86, 87, 89, 92, 94, and 98.

MARONNA, R. A.; ZAMAR, R. H. Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics*, Taylor & Francis, v. 44, n. 4, p. 307–317, 2002. Cited in page 95.

MARSILLI, C. Variable selection in predictive midas models. 2014. Cited in page 52.

MARTINS, A. C. Non-stationary correlation matrices and noise. *Physica A: Statistical Mechanics and its Applications*, Elsevier, v. 379, n. 2, p. 552–558, 2007. Cited 2 times in pages 43 and 88.

MARZAGÃO, T. *CATMATfinder classifier based on SVM*. 2015. https://github.com/thiagomarzagao/catmatfinder. Cited in page 33.

MASSARA, G. P.; MATTEO, T. D.; ASTE, T. Network filtering for big data: triangulated maximally filtered graph. *Journal of complex Networks*, Oxford University Press, v. 5, n. 2, p. 161–178, 2016. Cited 2 times in pages 40 and 90.

MCNEIL, A. J.; FREY, R. Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. *Journal of empirical finance*, Elsevier, v. 7, n. 3, p. 271–300, 2000. Cited in page 80.

MERTON, R. C. On estimating the expected return on the market: An exploratory investigation. *Journal of financial economics*, Elsevier, v. 8, n. 4, p. 323–361, 1980. Cited in page 53.

MILLER, M. H. The history of finance. *Journal of Portfolio Management*, INSTITUTIONAL INVESTOR INC, v. 25, p. 95–101, 1999. Cited in page 83.

MOGHADDAM, A. H.; MOGHADDAM, M. H.; ESFANDYARI, M. Stock market index prediction using artificial neural network. *Journal of Economics, Finance and Administrative Science*, Elsevier, v. 21, n. 41, p. 89–93, 2016. Cited in page 120.

MONTENEGRO, M. R.; ALBUQUERQUE, P. H. M. Wealth management: Modeling the nonlinear dependence. *Algorithmic Finance*, IOS Press, v. 6, n. 1-2, p. 51–65, 2017. Cited in page 89.

MOSSIN, J. Equilibrium in a capital asset market. *Econometrica: Journal of the econometric society*, JSTOR, p. 768–783, 1966. Cited in page 85.

MULLAINATHAN, S.; SPIESS, J. Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, v. 31, n. 2, p. 87–106, 2017. Cited in page 119.

MÜNNIX, M. C. et al. Identifying states of a financial market. *Scientific reports*, Nature Publishing Group, v. 2, p. 644, 2012. Cited in page 43.

MUSMECI, N.; ASTE, T.; MATTEO, T. D. What does past correlation structure tell us about the future? an answer from network filtering. *arXiv preprint arXiv:1605.08908*, 2016. Cited 2 times in pages 43 and 89.

NAKAMOTO, S. Bitcoin: A peer-to-peer electronic cash system. 2008. Cited in page 56.

NAKANO, M.; TAKAHASHI, A.; TAKAHASHI, S. Bitcoin technical trading with artificial neural network. *Physica A: Statistical Mechanics and its Applications*, v. 510, p. 587 – 609, 2018. Cited 6 times in pages 41, 123, 125, 127, 133, and 161.

NAYAK, A.; PAI, M. M.; PAI, R. M. Prediction models for indian stock market. *Procedia Computer Science*, Elsevier, v. 89, p. 441–449, 2016. Cited in page 120.

NAZÁRIO, R. T. F. et al. A literature review of technical analysis on stock markets. *The Quarterly Review of Economics and Finance*, Elsevier, v. 66, p. 115–126, 2017. Cited 2 times in pages 122 and 133.

NELSON, D. B. Conditional heteroskedasticity in asset returns: A new approach. *Econometrica: Journal of the Econometric Society*, JSTOR, p. 347–370, 1991. Cited in page 65.

NELSON, D. B. Filtering and forecasting with misspecified ARCH models I: Getting the right variance with the wrong model. *Journal of Econometrics*, Elsevier, v. 52, n. 1-2, p. 61–90, 1992. Cited in page 53.

NOBI, A. et al. Random matrix theory and cross-correlations in global financial indices and local stock market indices. *Journal of the Korean Physical Society*, Springer, v. 62, n. 4, p. 569–574, 2013. Cited 2 times in pages 91 and 118.

NOVAK, M. G.; VELUŠČEK, D. Prediction of stock price movement based on daily high prices. *Quantitative Finance*, Taylor & Francis, v. 16, n. 5, p. 793–826, 2016. Cited 3 times in pages 125, 126, and 127.

OLIVEIRA, F. A. de; NOBRE, C. N.; ZÁRATE, L. E. Applying artificial neural networks to prediction of stock price and improvement of the directional prediction index–case study of petr4, petrobras, brazil. *Expert Systems with Applications*, Elsevier, v. 40, n. 18, p. 7596–7606, 2013. Cited 4 times in pages 118, 125, 126, and 127.

OLIVEIRA, M. F. F. de. *Análise de mercado: uma ferramenta de mapeamento de oportunidades de negócio em técnicas de Geomarketing e Aprendizado de Máquina*. 2016. Term Paper, University of Brasilia. Cited in page 35.

PADULA, A. J. A. et al. *SEGURANÇA PÚBLICA E INTELIGÊNCIA ARTIFICIAL: UM ESTUDO GEORREFERENCIADO PARA O DISTRITO FEDERAL*. [S.l.], 2017. Cited in page 35.

PAIVA, F. D. et al. Decision-making for financial trading: A fusion approach of machine learning and portfolio selection. *Expert Systems with Applications*, Elsevier, 2018. Cited in page 88.

PAREEK, M. K.; THAKKAR, P. Surveying stock market portfolio optimization techniques. In: IEEE. *Engineering (NUiCONE), 2015 5th Nirma University International Conference on*. [S.l.], 2015. p. 1–5. Cited 2 times in pages 43 and 88.

PARK, J. et al. Some observations for portfolio management applications of modern machine learning methods. *International Journal of Fuzzy Logic and Intelligent Systems*, Korean Institute of Intelligent Systems, v. 16, n. 1, p. 44–51, 2016. Cited 2 times in pages 43 and 88.

PATEL, J. et al. Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert Systems with Applications*, Elsevier, v. 42, n. 1, p. 259–268, 2015. Cited 2 times in pages 125 and 126.

PATEL, J. et al. Predicting stock market index using fusion of machine learning techniques. *Expert Systems with Applications*, Elsevier, v. 42, n. 4, p. 2162–2172, 2015. Cited 3 times in pages 123, 125, and 126.

PATTON, A. J.; SHEPPARD, K. Good volatility, bad volatility: Signed jumps and the persistence of volatility. *Review of Economics and Statistics*, MIT Press, v. 97, n. 3, p. 683–697, 2015. Cited in page 101.

PAVLIDIS, E. G.; PAYA, I.; PEEL, D. A. Testing for linear and nonlinear granger causality in the real exchange rate–consumption relation. *Economics Letters*, Elsevier, v. 132, p. 13–17, 2015. Cited in page 84.

PEARSON, K. Principal components analysis. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, v. 6, n. 2, p. 559, 1901. Cited in page 95.

PETROPOULOS, A. et al. A stacked generalization system for automated forex portfolio trading. *Expert Systems with Applications*, Elsevier, v. 90, p. 290–302, 2017. Cited in page 88.

PLEROU, V. et al. Random matrix approach to cross correlations in financial data. *Physical Review E*, APS, v. 65, n. 6, p. 066126, 2002. Cited in page 91.

POLASIK, M. et al. Price fluctuations and the use of bitcoin: An empirical inquiry. *International Journal of Electronic Commerce*, Taylor & Francis, v. 20, n. 1, p. 9–49, 2015. Cited in page 57.

POON, S.-H.; GRANGER, C. Forecasting volatility in financial markets: A review. *Journal of economic literature*, American Economic Association, v. 41, n. 2, p. 478–539, 2003. Cited in page 53.

PREMANODE, B.; TOUMAZOU, C. Improving prediction of exchange rates using differential emd. *Expert systems with applications*, Elsevier, v. 40, n. 1, p. 377–384, 2013. Cited 2 times in pages 62 and 66.

PUDIL, P.; NOVOVIČOVÁ, J.; KITTLER, J. Floating search methods in feature selection. *Pattern recognition letters*, Elsevier, v. 15, n. 11, p. 1119–1125, 1994. Cited in page 134.

PéRIGNON, C.; SMITH, D. R.; VILLA, C. Why common factors in international bond returns are not so common. *Journal of International Money and Finance*, v. 26, n. 2, p. 284 – 304, 2007. ISSN 0261-5606. Disponível em: <http://www.sciencedirect.com-/science/article/pii/S0261560606001331>. Cited in page 96.

QIU, M.; SONG, Y.; AKAGI, F. Application of artificial neural network for the prediction of stock market returns: The case of the japanese stock market. *Chaos, Solitons & Fractals*, Elsevier, v. 85, p. 1–7, 2016. Cited in page 120.

RAMAKRISHNAN, S. et al. Forecasting malaysian exchange rate using machine learning techniques based on commodities prices. In: IEEE. *Research and Innovation in Information Systems (ICRIIS), 2017 International Conference on*. [S.l.], 2017. p. 1–5. Cited in page 39.

RANJAN, J.; GOYAL, D.; AHSON, S. Data mining techniques for better decisions in human resource management systems. *International Journal of Business Information Systems*, Inderscience Publishers, v. 3, n. 5, p. 464–481, 2008. Cited in page 36.

REBOREDO, J. C.; MATÍAS, J. M.; GARCIA-RUBIO, R. Nonlinearity in forecasting of high-frequency stock returns. *Computational Economics*, Springer, v. 40, n. 3, p. 245–264, 2012. Cited 2 times in pages 49 and 63.

REN, F.; ZHOU, W.-X. Dynamic evolution of cross-correlations in the chinese stock market. *PloS one*, Public Library of Science, v. 9, n. 5, p. e97711, 2014. Cited in page 92.

REUNANEN, J. Overfitting in making comparisons between variable selection methods. *Journal of Machine Learning Research*, v. 3, n. Mar, p. 1371–1382, 2003. Cited in page 135.

RODRÍGUEZ-GONZÁLEZ, A. et al. Cast: Using neural networks to improve trading systems based on technical analysis by means of the rsi financial indicator. *Expert systems with Applications*, Elsevier, v. 38, n. 9, p. 11489–11500, 2011. Cited in page 125.

ROM, B. M.; FERGUSON, K. W. Post-modern portfolio theory comes of age. *Journal of Investing*, v. 3, n. 3, p. 11–17, 1994. Cited in page 86.

ROSENFELD, M. Analysis of hashrate-based double spending. *arXiv preprint arXiv:1402.2009*, 2014. Cited in page 56.

ROUSSEEUW, P. J. Least median of squares regression. *Journal of the American statistical association*, Taylor & Francis, v. 79, n. 388, p. 871–880, 1984. Cited in page 95.

ROUSSEEUW, P. J.; DRIESSEN, K. V. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, Taylor & Francis Group, v. 41, n. 3, p. 212–223, 1999. Cited in page 95.

RUNGE, C. Über empirische funktionen und die interpolation zwischen äquidistanten ordinaten. *Zeitschrift für Mathematik und Physik*, v. 46, n. 224-243, p. 20, 1901. Cited in page 46.

SAAVEDRA, C. A. P. B. *Análise de Ampla Associação do Genoma: Redução de Dimensionalidade via Seleção por Torneios e Análises Indicativas*. 2015. Term Paper, University of Brasilia. Cited in page 136.

SAGIROGLU, S.; SINANC, D. Big data: A review. In: IEEE. *2013 International Conference on Collaboration Technologies and Systems (CTS)*. [S.l.], 2013. p. 42–47. Cited in page 30.

SALCEDO-SANZ, S. et al. Feature selection methods involving support vector machines for prediction of insolvency in non-life insurance companies. *Intelligent Systems in Accounting, Finance & Management: International Journal*, Wiley Online Library, v. 12, n. 4, p. 261–281, 2004. Cited in page 118.

SANDOVAL-JR, L.; BORTOLUZZO, A. B.; VENEZUELA, M. K. Not all that glitters is rmt in the forecasting of risk of portfolios in the brazilian stock market. *Physica A: Statistical Mechanics and its Applications*, Elsevier, v. 410, p. 94–109, 2014. Cited 2 times in pages 111 and 112.

SANTAMARÍA-BONFIL, G.; FRAUSTO-SOLÍS, J.; VÁZQUEZ-RODARTE, I. Volatility forecasting using support vector regression and a hybrid genetic algorithm. *Computational Economics*, Springer, v. 45, n. 1, p. 111–133, 2015. Cited 3 times in pages 62, 66, and 68.

SANTOS, A. A. P.; COSTA, N. C. A. da; COELHO, L. dos S. Computational intelligence approaches and linear models in case studies of forecasting exchange rates. *Expert Systems with Applications*, Elsevier, v. 33, n. 4, p. 816–823, 2007. Cited in page 66.

SCHÖLKOPF, B.; SMOLA, A. J. *Learning with kernels: Support Vector Machines, Regularization, Optimization adn Beyond*. [S.l.]: The MIT Press, 2002. Cited in page 39.

SCHÜRITZ, R.; SATZGER, G. Patterns of data-infused business model innovation. In: IEEE. *2016 IEEE 18th Conference on Business Informatics (CBI)*. [S.l.], 2016. v. 1, p. 133–142. Cited in page 38.

SCIKIT-LEARN. *Scikit-Learn documentation, version 0.19.1*. 2017. http://scikit-learn.org/stable/auto_examples/model_selection/plot_underfitting_overfitting.html. Accessed: 04-06-2018. Cited in page 47.

SEGENDORF, B. Have virtual currencies affected the retail payments market. *Sveriges Riksbank Economic Commentaries*, v. 2, p. 1–5, 2014. Cited in page 57.

SENSOY, A.; YUKSEL, S.; ERTURK, M. Analysis of cross-correlations between financial markets after the 2008 crisis. *Physica A: Statistical Mechanics and its Applications*, Elsevier, v. 392, n. 20, p. 5027–5045, 2013. Cited 3 times in pages 91, 112, and 118.

SERMPINIS, G. et al. Modeling, forecasting and trading the eur exchange rates with hybrid rolling genetic algorithms—support vector regression forecast combinations. *European Journal of Operational Research*, Elsevier, v. 247, n. 3, p. 831–846, 2015. Cited in page 62.

SEWELL, M. V. *Application of Machine Learning to Financial Time Series Analysis*. Tese (Doutorado) — UCL (University College London), 2017. Cited in page 41.

SHARIFI, S. et al. Random matrix theory for portfolio optimization: A stability approach. *Physica A: Statistical Mechanics and its Applications*, v. 335, n. 3-4, p. 629–643, 2004. ISSN 03784371. Cited in page 98.

SHARMA, C.; BANERJEE, K. A study of correlations in the stock market. *Physica A: Statistical Mechanics and its Applications*, Elsevier, v. 432, p. 321–330, 2015. Cited in page 92.

SHARPE, W. F. Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance*, Wiley Online Library, v. 19, n. 3, p. 425–442, 1964. Cited in page 85.

SHARPE, W. F. Mutual Fund Performance. *The Journal of Business*, v. 39, n. S1, p. 119, 1966. ISSN 0021-9398. Disponível em: <http://www.jstor.org/stable/2351741>. Cited in page 101.

SHEKAR, B. et al. Face recognition using kernel entropy component analysis. *Neurocomputing*, v. 74, n. 6, p. 1053 – 1057, 2011. ISSN 0925-2312. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0925231210004881>. Cited in page 97.

SHEN, F.; CHAO, J.; ZHAO, J. Forecasting exchange rate using deep belief networks and conjugate gradient method. *Neurocomputing*, Elsevier, v. 167, p. 243–253, 2015. Cited 2 times in pages 62 and 66.

SIVARAJAH, U. et al. Critical analysis of big data challenges and analytical methods. *Journal of Business Research*, Elsevier, v. 70, p. 263–286, 2017. Cited in page 29.

SORESCU, A. Data-driven business model innovation. *Journal of Product Innovation Management*, Wiley Online Library, v. 34, n. 5, p. 691–696, 2017. Cited in page 38.

SORTINO, F. A.; PRICE, L. N. Performance measurement in a downside risk framework. *The Journal of Investing*, Institutional Investor Journals Umbrella, v. 3, n. 3, p. 59–64, 1994. Cited in page 101.

SPAGNOLO, B.; VALENTI, D. Volatility effects on the escape time in financial market models. *International Journal of Bifurcation and Chaos*, World Scientific, v. 18, n. 09, p. 2775–2786, 2008. Cited in page 87.

SRIVASTAVA, N. et al. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, JMLR. org, v. 15, n. 1, p. 1929–1958, 2014. Cited in page 133.

StockCharts. *Technical Indicators and Overlays*. 2019. https://stockcharts.com/school/doku.php?id=chart_school:technical_indicators. Cited 3 times in pages 128, 129, and 130.

SUN, P.; ZHOU, C. Diagnosing the distribution of garch innovations. *Journal of Empirical Finance*, Elsevier, v. 29, p. 287–303, 2014. Cited in page 65.

TAY, F. E.; CAO, L. Application of support vector machines in financial time series forecasting. *omega*, Elsevier, v. 29, n. 4, p. 309–317, 2001. Cited 3 times in pages 125, 126, and 127.

TAYALI, H. A.; TOLUN, S. Dimension reduction in mean-variance portfolio optimization. *Expert Systems with Applications*, Elsevier, v. 92, p. 161–169, 2018. Cited in page 89.

THAWORNWONG, S.; ENKE, D.; DAGLI, C. Neural networks as a decision maker for stock trading: a technical analysis approach. *International Journal of Smart Engineering System Design*, Taylor & Francis, v. 5, n. 4, p. 313–325, 2003. Cited 2 times in pages 125 and 127.

TIAN, F.; YANG, K.; CHEN, L. Realized volatility forecasting of agricultural commodity futures using the har model with time-varying sparsity. *International Journal of Forecasting*, Elsevier, v. 33, n. 1, p. 132–152, 2017. Cited in page 51.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, JSTOR, p. 267–288, 1996. Cited in page 136.

TICKNOR, J. L. A bayesian regularized artificial neural network for stock market forecasting. *Expert Systems with Applications*, Elsevier, v. 40, n. 14, p. 5501–5506, 2013. Cited 2 times in pages 125 and 126.

TITIUNIK, R. Can big data solve the fundamental problem of causal inference? *PS: Political Science & Politics*, Cambridge University Press, v. 48, n. 1, p. 75–79, 2015. Cited in page 31.

TORUN, M. U.; AKANSU, A. N.; AVELLANEDA, M. Portfolio risk in multiple frequencies. *IEEE Signal Processing Magazine*, IEEE, v. 28, n. 5, p. 61–71, 2011. Cited 2 times in pages 48 and 90.

TRACY, C. A.; WIDOM, H. Distribution functions for largest eigenvalues and their applications. *arXiv preprint math-ph/0210034*, 2002. Cited in page 92.

Trading Technologies. *X_STUDY List of Technical Indicators*. 2019. www.tradingtechnologies.com/help/x-study/technical-indicator-definitions/list-of-technical-indicators. Cited 3 times in pages 128, 129, and 130.

TradingView. *Indicators and Overlays*. 2019. https://www.tradingview.com/wiki/Category:Indicators_and_overlays. Cited 3 times in pages 128, 129, and 130.

TREYNOR, J. L.; BLACK, F. How to use security analysis to improve portfolio selection. *The Journal of Business*, JSTOR, v. 46, n. 1, p. 66–86, 1973. Cited in page 86.

TYLER, D. E. Robustness and Efficiency Properties of Scatter Matrices 2. *Biometrika*, v. 70, n. 2, p. 411–420, 1983. Cited in page 98.

URQUHART, A. The inefficiency of bitcoin. *Economics Letters*, Elsevier, v. 148, p. 80–82, 2016. Cited in page 59.

VALENTI, D.; FAZIO, G.; SPAGNOLO, B. Stabilizing effect of volatility in financial markets. *Physical Review E*, APS, v. 97, n. 6, p. 062307, 2018. Cited in page 87.

VANSTONE, B.; FINNIE, G. Enhancing stockmarket trading performance with anns. *Expert Systems with Applications*, Elsevier, v. 37, n. 9, p. 6602–6610, 2010. Cited 3 times in pages 125, 126, and 127.

VAPNIK, V.; LEVIN, E.; CUN, Y. L. Measuring the vc-dimension of a learning machine. *Neural computation*, MIT Press, v. 6, n. 5, p. 851–876, 1994. Cited in page 45.

VAPNIK, V. N. *The nature of statical learning theory*. [S.l.]: Springer Verlag, 1995. Cited in page 66.

VARIAN, H. R. Big data: New tricks for econometrics. *Journal of Economic Perspectives*, v. 28, n. 2, p. 3–28, 2014. Cited in page 39.

VASQUEZ, A. Venezuelans use bitcoin 'mining' to escape inflation. *Yahoo! News*, 2017. Online; posted 23-October-2017. Disponível em: <https://www.yahoo.com-/news/venezuelans-bitcoin-mining-escape-inflation-020507653.html>. Cited in page 56.

VELOSO, E. F. R. et al. The use of traditional and non-traditional career theories to understand the young's relationship with new technologies. *Revista de Gestão*, Emerald Publishing Limited, v. 25, n. 4, p. 340–357, 2018. Cited in page 36.

VIGNA, P.; CASEY, M. J. *The age of cryptocurrency: how bitcoin and the blockchain are challenging the global economic order*. [S.l.]: Macmillan, 2016. Cited in page 55.

WALLER, M. A.; FAWCETT, S. E. Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management. *Journal of Business Logistics*, Wiley Online Library, v. 34, n. 2, p. 77–84, 2013. Cited in page 35.

WAMBA, S. F. et al. How 'big data'can make big impact: Findings from a systematic review and a longitudinal case study. *International Journal of Production Economics*, Elsevier, v. 165, p. 234–246, 2015. Cited in page 30.

WAN, X.; PEKNY, J.; REKLAITIS, G. Simulation based optimization for risk management in multi-stage capacity expansion. In: *Computer Aided Chemical Engineering*. [S.l.]: Elsevier, 2006. v. 21, p. 1881–1886. Cited in page 93.

WANG, F.-Y. et al. Societies 5.0: A new paradigm for computational social systems research. *IEEE Transactions on Computational Social Systems*, IEEE, v. 5, n. 1, p. 2–8, 2018. Cited in page 31.

WANG, G. et al. Big data analytics in logistics and supply chain management: Certain investigations for research and applications. *International Journal of Production Economics*, Elsevier, v. 176, p. 98–110, 2016. Cited in page 35.

WANG, Y. et al. Forecasting realized volatility in a changing world: A dynamic model averaging approach. *Journal of Banking & Finance*, Elsevier, v. 64, p. 136–149, 2016. Cited in page 51.

WANG, Y.-H. Nonlinear neural network forecasting model for stock index option price: Hybrid gjr–garch approach. *Expert Systems with Applications*, Elsevier, v. 36, n. 1, p. 564–570, 2009. Cited in page 65.

WARING, E. Vii. problems concerning interpolations. *Philosophical transactions of the royal society of London*, The Royal Society London, n. 69, p. 59–67, 1779. Cited in page 46.

WENG, B.; AHMED, M. A.; MEGAHED, F. M. Stock market one-day ahead movement prediction using disparate data sources. *Expert Systems with Applications*, Elsevier, v. 79, p. 153–163, 2017. Cited 3 times in pages 123, 125, and 126.

WHITE, M. Digital workplaces: Vision and reality. *Business information review*, Sage Publications Sage UK: London, England, v. 29, n. 4, p. 205–214, 2012. Cited in page 29.

WOLPERT, D. H.; MACREADY, W. G. No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, IEEE, v. 1, n. 1, p. 67–82, 1997. Cited in page 44.

XIE, H.; LI, J. Intraday volatility analysis on s&p 500 stock index future. *International Journal of Economics and Finance*, v. 2, n. 2, p. 26, 2010. Cited in page 78.

XIONG, T.; BAO, Y.; HU, Z. Multiple-output support vector regression with a firefly algorithm for interval-valued stock price index forecasting. *Knowledge-Based Systems*, Elsevier, v. 55, p. 87–100, 2014. Cited in page 59.

YAOHAO, P.; ALBUQUERQUE, P. H. M. Non-linear interactions and exchange rate prediction: Empirical evidence using support vector regression. *Applied Mathematical Finance*, Taylor & Francis, p. 1–32, 2019. Cited 2 times in pages 44 and 162.

YERMACK, D. *Is Bitcoin a real currency? An economic appraisal*. [S.l.], 2013. Cited 4 times in pages 54, 58, 59, and 60.

YU, L. et al. Evolving least squares support vector machines for stock market trend mining. *IEEE Transactions on evolutionary computation*, IEEE, v. 13, n. 1, p. 87–102, 2009. Cited 3 times in pages 125, 126, and 127.

ZAKOIAN, J.-M. Threshold heteroskedastic models. *Journal of Economic Dynamics and control*, Elsevier, v. 18, n. 5, p. 931–955, 1994. Cited in page 80.

ŻBIKOWSKI, K. Using volume weighted support vector machines with walk forward testing and feature selection for the purpose of creating stock trading strategy. *Expert Systems with Applications*, Elsevier, v. 42, n. 4, p. 1797–1805, 2015. Cited 2 times in pages 122 and 123.

ZHANG, W.-G.; ZHANG, X.-L.; XIAO, W.-L. Portfolio selection under possibilistic mean–variance utility and a smo algorithm. *European Journal of Operational Research*, Elsevier, v. 197, n. 2, p. 693–700, 2009. Cited in page 86.

ZHENG, Z. et al. Changes in cross-correlations as an indicator for systemic risk. *Scientific reports*, Nature Publishing Group, v. 2, p. 888, 2012. Cited in page 96.

ZHONG, R. Y. et al. A big data approach for logistics trajectory discovery from rfid-enabled production data. *International Journal of Production Economics*, Elsevier, v. 165, p. 260–272, 2015. Cited in page 36.

ZHU, H. et al. Profitability of simple technical trading rules of chinese stock exchange indexes. *Physica A: Statistical Mechanics and its Applications*, Elsevier, v. 439, p. 75–84, 2015. Cited in page 123.

ZHU, Z.; WELSCH, R. E. et al. Robust dependence modeling for high-dimensional covariance matrices with financial applications. *The Annals of Applied Statistics*, Institute of Mathematical Statistics, v. 12, n. 2, p. 1228–1249, 2018. Cited in page 95.