UNIVERSITY OF THE AEGEAN

DOCTORAL THESIS

Recent Advances on Dimensionality Reduction for High-dimensional Data Analysis with Applications

Author:
Kimon NTOTSIS

Supervisor:
Prof. Alex KARAGRIGORIOU

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

in the

Department of Statistics and Actuarial-Financial Mathematics
School of Sciences

October 17, 2022

Με ατομική μου ευθύνη και γνωρίζοντας τις κυρώσεις, που προβλέπονται από τις διατάξεις της παρ. 6 του άρθρου 22 του Ν. 1599/1986, δηλώνω ότι:

Είμαι ο αποκλειστικός συγγραφέας της υποβληθείσας Διδακτορικής Διατριβής με τίτλο «Recent Advances on Dimensionality Reduction for High-dimensional Data Analysis with Applications». Η συγκεκριμένη Διδακτορική Διατριβή είναι πρωτότυπη και εκπονήθηκε αποκλειστικά για την απόκτηση του Διδακτορικού διπλώματος του Τμήματος. Κάθε βοήθεια, την οποία είχα για την προετοιμασία της, αναγνωρίζεται πλήρως και αναφέρεται επακριβώς στην εργασία. Επίσης, επακριβώς αναφέρω στην εργασία τις πηγές, τις οποίες χρησιμοποίησα, και μνημονεύω επώνυμα τα δεδομένα ή τις ιδέες που αποτελούν προϊόν πνευματικής ιδιοκτησίας άλλων, ακόμη κι εάν η συμπερίληψή τους στην παρούσα εργασία υπήρξε έμμεση ή παραφρασμένη. Γενικότερα, βεβαιώνω ότι κατά την εκπόνηση της Διδακτορικής Διατριβής έχω τηρήσει απαρέγκλιτα όσα ο νόμος ορίζει περί διανοητικής ιδιοκτησίας και έχω συμμορφωθεί πλήρως με τα προβλεπόμενα στο νόμο περί προστασίας προσωπικών δεδομένων και τις αρχές Ακαδημαϊκής Δεοντολογίας.

UNIVERSITY OF THE AEGEAN

DOCTORAL THESIS

Recent Advances on Dimensionality Reduction for High-dimensional Data Analysis with Applications

Author:
Kimon NTOTSIS

Supervisor:
Prof. Alex KARAGRIGORIOU

Seven-member Evaluation Committee:

ARTEMIOU Andreas (*3-member Advisory Committee*)
Reader in Statistics, Cardiff University

GAKI Eleni
Assistant Professor, University of the Aegean

HATZOPOULOS Petros
Associate Professor, University of the Aegean

KARAGRIGORIOU Alex (*supervisor*)
Professor, University of the Aegean

KOUNTZAKIS Christos
Assistant Professor University of the Aegean

RAKITZIS Athanasios (*3-member Advisory Committee*)
Assistant Professor, University of Piraeus

TSILIKA Kyriaki
Associate Professor, University of Thessaly

UNIVERSITY OF THE AEGEAN

Abstract

School of Sciences
Department of Statistics and Actuarial-Financial Mathematics

Doctor of Philosophy

Recent Advances on Dimensionality Reduction for High-dimensional Data Analysis with Applications

by Kimon NTOTSIS

Large amounts of raw data often can fail to perform properly for model estimation, attributed to the existence of multicollinearity between variables, and that is why they must be pre-processed for better modeling and visualization. To address raw data barriers, among other difficulties, Dimension Reduction Techniques were developed in an effort to mitigate the magnitude of over-parametrized solutions that arise in high-dimensional spaces. The aim of this dissertation, which utilizes multivariate analysis tools, is to investigate, analyze, compare, and improve current techniques while still introducing new ones for dealing with multicollinearity and reducing the feature space of high-dimensional data. In particular, this doctoral thesis initially outlines the theoretical framework concerning the unsupervised technique, Principal Component Analysis, and its supervised counterpart, the Partial Least Squares method. Due to their ability to obtain dimension reduction when analyzing high-dimensional datasets, both techniques are considered optimal for feature extraction. The use of the former in conjunction with other dimension reduction techniques, as well as the modification of the latter, - so that it may be applied as a feature selection and feature extraction simultaneously-, were implemented and thoroughly studied in the fields of econometrics, finance and actuarial science. Finally, a new unsupervised linear feature selection technique is proposed as a robust and easily interpretable methodology, termed Elastic Information Criterion, that is capable of capturing multicollinearity rather accurately and effectively and thus providing a proper dataset assessment.

ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

Περίληψη

Σχολή Θετικών Επιστημών
Τμήμα Στατιστικής και Αναλογιστικών-Χρηματοοικονομικών Μαθηματικών

Διδακτορικό

Πρόσφατες Εξελίξεις για Μείωση Διάστασης στην Πολυδιάστατη Ανάλυση Δεδομένων με Εφαρμογές

του Κίμων ΝΤΟΤΣΗ

Η ιλιγγιώδης ανάπτυξη της τεχνολογίας καθώς και η συνεχής δημιουργία νέων βάσεων δεδομένων έχει αναδείξει και αναγάγει τα *Big Data Analytics* στο επίκεντρο της Σύγχρονης Στατιστικής Ανάλυσης δεδομένων. Ωστόσο, η συνεχής αύξηση των παρατηρήσεων σε ένα σετ δεδομένων δεν προϋποθέτει πάντα μια πιο αποδοτική στατιστική ανάλυση. Μεγάλοι όγκοι δεδομένων ενδέχεται να παρεμποδίζουν την αποδοτικότητα και αποτελεσματικότητα των διαφόρων στατιστικών τεχνικών. Συγχρόνως , στην περίπτωση της μοντελοποίησης, μια πληθώρα επεξηγηματικών μεταβλητών για εκτεταμένες χρονικές περιόδους μπορεί να προκαλέσει ασυνέπειες στην ερμηνεία των στατιστικών αποτελεσμάτων. Το πιο σημαντικό εμπόδιο που καλείται να προσπελάσει ο ερευνητής είναι η ύπαρξη πολυσυγγραμμικότητας μεταξύ των μεταβλητών. Πλέον οι ερευνητές που εμπλέκονται σε διαδικασίες στατιστικής μοντελοποίησης, είναι σε θέση να συλλέγουν και να δημιουργούν σετ μεγάλων πολυμεταβλητών δεδομένων, δηλαδή σετ που περιέχουν πολλές μεταβλητές με πάρα πολλές παρατηρήσεις ή/και καλύπτουν μεγάλο χρονικό ορίζοντα. Η προσπάθεια μοντελοποίησης μιας μεταβλητής με τη χρήση ενός τέτοιου σετ ενδέχεται να αναδείξει επιπρόσθετα προβλήματα -όπως ανακριβείς και αδόμητες βάσεις δεδομένων καθώς και υπολογιστική πολυπλοκότητα, σε αντίθεση με τη μοντελοποίηση με τη χρήση μικρότερων συνόλων.

Η πιο συνήθης συνέπεια στη μοντελοποίηση μεγάλων δεδομένων στη στατιστική ανάλυση είναι η ύπαρξη πολυσυγγραμμικότητας, η οποία ορίζεται ως η υψηλή γραμμική συσχέτιση μεταξύ δύο ή περισσότερων μεταβλητών. Σε μικρότερη συχνότητα εντοπίζεται ότι η πηγή του προβλήματος μπορεί να οφείλεται σε εσφαλμένη χρήση ψευδομεταβλητών στη μοντελοποίηση, στην ύπαρξη μεταβλητών που εκφράζουν το ίδιο χαρακτηριστικό και έχουν οριστεί με διαφορετικό τρόπο -άλλη μονάδα μέτρησης ή κατηγοριοποίησης- ή όταν τα δεδομένα είναι ανεπαρκή. Η συνύπαρξη αυτών των «συγγενών» μεταβλητών παραδείγματος χάριν στην ανάλυση παλινδρόμησης μπορεί να οδηγήσει σε ασαφή ή και λανθασμένη ερμηνεία ενώ συγχρόνως μπορεί να επηρεάσει τη διαδικασία πρόβλεψης. Παρόλο που η σχετικά μικρή πολυσυγγραμμικότητα είναι συνήθως αβλαβής, η μέτρια και σοβαρή ενδέχεται να βλάψει τη στατιστική ισχύ της παλινδρόμησης και να οδηγήσει σε υπερπροσαρμογή του μοντέλου. Αυτό το φαινόμενο είναι αρκετά συχνό στις μέρες μας λόγω του πλήθους των μη ελεγχόμενων πληροφοριών. Η έλλειψη «φίλτρων» στα δεδομένα οδηγεί σε σύνολα δεδομένων με μεταβλητές που είναι, σε σημαντικό βαθμό, συγγραμμικές λόγω αλληλεπιδράσεων που υποβόσκουν και δυνητικά οδηγούν σε παραπλανητικά μοντέλα (υπερεκτίμηση/υποεκτίμηση) και ανακρίβεια στην εκτίμηση παραμέτρων.

Όσο υφίσταται ο κίνδυνος της πολυσυγγραμμικότητας, η ορθή ερμηνεία της ανάλυσης δεδομένων μπορεί να χαρακτηριστεί ως αναξιόπιστη. Ως εκ τούτου, η πολυσυγγραμμικότητα έχει φτάσει στο σημείο να αποτελεί μια επιπρόσθετη «πέμπτη» προϋπόθεση στην

περίπτωση της πολλαπλής και πολυμεταβλητής παλινδρόμησης, μεταξύ των ήδη υπαρχόντων που πρέπει να ελεγχθούν πριν την διαδικασία μοντελοποίησης.

Η πολυσυγγραμμικότητα μπορεί να χαρακτηρισθεί ως «δομική» ή «βασισμένη στα δεδομένα». Η δομική, γνωστή και ως τέλεια πολυσυγγραμμικότητα, εμφανίζεται όταν σε ένα σύνολο δεδομένων υπάρχουν μεταβλητές οπού η ύπαρξη τους είναι άρρηκτα συνδεδεμένη με κάποια άλλη μεταβλητή και δεν μπορούν να οριστούν χωρίς εκείνη. Η συσχέτιση αυτών των δύο μεταβλητών είναι προσεγγιστικά γραμμική και μπορεί να περιγράφει (υποθέτοντας ότι υπάρχει συσχέτιση μεταξύ τουλάχιστον δύο μεταβλητών απο τις συνολικά p ενός πίνακα X) ως:

$$\sum_{j=1}^{p} a_j X_j = 0$$

όπου υπάρχουν $a_j \neq 0$ που επιβεβαιώνουν την εξίσωση. Με άλλα λόγια, είναι ένα μαθηματικό τέχνασμα που προκαλείται από τη δημιουργία μεταβλητών με τη χρήση ήδη υπαρχόντων μεταβλητών από το ίδιο σύνολο δεδομένων. Λόγω του ότι υπάρχει μια τέλεια συγγραμμική σχέση μεταξύ των μεταβλητών που περιλαμβάνονται στο μοντέλο, το $X^{\intercal}X$ γίνεται ιδιάζων και επομένως δεν είναι εφικτό να χρησιμοποιηθεί η παλινδρόμηση ελαχίστων τετραγώνων για την εκτίμηση των τιμών των παραμέτρων λόγω της μη αντιστρεψιμότητας του $X^{\intercal}X$. Επομένως, η τέλεια πολυσυγγραμμικότητα παραβιάζει μία από τις προϋποθέσεις του μοντέλου γραμμικής παλινδρόμησης.

Η βασιζόμενη στα δεδομένα, γνωστή και ως υψηλή πολυσυγγραμμικότητα, εμφανίζεται μεταξύ των μεταβλητών στο αρχικό μη επεξεργασμένο σύνολο δεδομένων και είναι ο πιο κοινός τύπος όταν πρόκειται για παρατηρήσεις πειραμάτων. Σε αυτήν την περίπτωση, η σχέση μεταξύ των μεταβλητών είναι προσεγγιστικά γραμμική και μπορεί να γραφτεί ως:

$$\sum_{j=1}^{p} a_j X_j \approx 0$$

όπου υπάρχουν $a_j \neq 0$ που επιβεβαιώνουν την εξίσωση. Υψηλή πολυσυγγραμμικότητα εντοπίζεται συχνά σε μεγάλα πολυμεταβλητά σύνθετα σύνολα δεδομένων, όπου οι μεταβλητές μπορούν να ποσοτικοποιηθούν σε μετρήσεις ανόμοιου μεγέθους, το οποίο ενέχει την ενίσχυση της σημαντικότητας των μη σημαντικών μεταβλητών και κατ' επέκταση την απόκρυψη της ισχύς των πραγματικά στατιστικά σημαντικών. Σε αυτήν την περίπτωση, μπορούν να προκύψουν οι ακόλουθες καταστάσεις σε μια διαδικασία μοντελοποίησης:
(i) Καθίσταται δύσκολη η διάκριση των επιδράσεων των επεξηγηματικών μεταβλητών στη μεταβλητή απόκρισης κατά τη διάρκεια της διαδικασίας της μοντελοποίησης. Όταν οι επεξηγηματικές μεταβλητές συσχετίζονται, μοιράζονται ένα κοινό μοτίβο μεταβλητότητας, γεγονός που σημαίνει ότι αυξάνονται και μειώνονται ταυτόχρονα. Ως αποτέλεσμα, η εκτίμηση της επίδρασης στη μεταβλητή απόκρισης είναι δύσκολη και ως εκ τούτου, τα παραγόμενα αποτελέσματα είναι παραπλανητικά.
(ii) Επιπλέον, στην παρουσίαση πολυσυγγραμμικότητας, υπάρχουν πολυάριθμοι συνδυασμοί εκτιμώμενων συντελεστών που όλοι παράγουν παρόμοιες προβλέψεις. Αυτό σημαίνει ότι οι εκτιμήσεις έχουν υψηλή τυπική απόκλιση. Σε αυτές τις περιπτώσεις, ο συντελεστής προσδιορισμού -το μέτρο που χρησιμοποιείται για την αξιολόγηση της καλής προσαρμογής ενός μοντέλου-, θα είναι υψηλός και το παραγόμενο μοντέλο θα ερμηνεύεται, εσφαλμένα λόγω υπερπροσαρμογής, ως επαρκές.

Η τέλεια πολυσυγγραμμικότητα είναι εξαιρετικά ασυνήθιστη και η πιο εύκολα διαχειρίσιμη αφού μια ενδελεχής εξέταση των μεταβλητών του μοντέλου μπορεί να την εξαλείψει.

Ωστόσο, η υψηλή πολυσυγγραμμικότητα -η πιο συνηθισμένη, μπορεί να προκαλέσει σοβαρά προβλήματα εκτίμησης και ερμηνείας. Η πιο διαδεδομένη συνέπεια που εμφανίζεται παρουσία υψηλής πολυσυγγραμμικότητας είναι η υπερπροσαρμογή της μοντελοποίησης λόγω της υπερπαραμετροποίησης που υφίσταται, γεγονός που μειώνει τη ισχύ του μοντέλου στο προσδιορισμό των στατιστικά σημαντικών μεταβλητών. Αυτό σημαίνει ότι το μοντέλο είναι πολύ περίπλοκο ενώ οι δείκτες αξιολόγησης του, όπως ο συντελεστής προσδιορισμού, είναι παραπλανητικοί. Αντί να περιγράφουν την αναλογία της διακύμανσης στην εξαρτημένη μεταβλητή που είναι προβλέψιμη από τις ανεξάρτητες μεταβλητές, περιγράφουν το τυχαίο σφάλμα στα δεδομένα. Είναι επίσης πιθανό οι εκτιμήσεις παραμέτρων να μην περιγράφουν με ακρίβεια την επίδραση των σχετικών μεταβλητών στην εξαρτημένη μεταβλητή. Μπορεί επίσης να έχει ως αποτέλεσμα την αλλαγή του δείκτη και την απεραντοσύνη των συντελεστών μερικής παλινδρόμησης από το ένα δείγμα στο άλλο. Επιπλέον, αν και αυτό το φαινόμενο φαίνεται αρκετά παράδοξο, έχουν γίνει αναφορές για μη συμβατά αποτελέσματα μεταξύ του F-test και του t-test όταν υπάρχει πολυσυγγραμμικότητα.

Η πολυσυγγραμικότητα ωστόσο δεν είναι πάντα επιβλαβής. Υπάρχουν περιπτώσεις που μπορεί να χαρακτηριστεί ως ανύπαρκτη, χαμηλή ή μέτρια. Σε αυτές τις περιπτώσεις δεν χρειάζεται καμία διαδικασία για την αντιμετώπιση της.

Για την αντιμετώπιση της τροχοπέδης που δημιουργεί η ύπαρξη πολυσυγγραμμικοτητας αναπτύχθηκαν οι λεγόμενες Τεχνικές Μείωσης Διάστασης σε μια προσπάθεια να μετριαστεί το μέγεθος των υπερβολικά παραμετροποιημένων λύσεων που προκύπτουν σε χώρους υψηλών διαστάσεων. Στόχος της παρούσας διατριβής, η οποία χρησιμοποιεί εργαλεία πολυμεταβλητής ανάλυσης, είναι να διερευνήσει, να αναλύσει, να συγκρίνει και να βελτιώσει ήδη υπάρχουσες τεχνικές, ενώ παράλληλα να εισάγει νέες για την αντιμετώπιση της πολυσυγγραμμικότητας και τη μείωση του χώρου χαρακτηριστικών των δεδομένων υψηλών διαστάσεων.

Πιο συγκεκριμένα στο Κεφάλαιο 2 σκιαγραφείται το θεωρητικό πλαίσιο που αφορά την μη εποπτευόμενη τεχνική της Ανάλυσης Κύριων Συνιστωσών (ΑΚΣ) καθώς και την αντίστοιχη εποπτευόμενη μέθοδο των Μερικών Ελάχιστων Τετράγωνων (ΜΕΤ). Λόγω της ικανότητάς τους να επιτυγχάνουν μείωση διαστάσεων κατά την ανάλυση συνόλων δεδομένων υψηλών διαστάσεων, θεωρούνται αμφότερες βέλτιστες για δημιουργία νέων μεταβλητών.

Στο Κεφάλαιο 3 γίνεται αναφορά στα κριτήρια και στους δείκτες που έχουν χρησιμοποιηθεί για την αξιολόγηση των τεχνικών που εφαρμόστηκαν στη διεκπεραίωση της διατριβής.

Ο σκοπός του Κεφαλαίου 4 αφορά τη μοντελοποίηση των δημοσίων συνταξιοδοτικών δαπανών (ΔΣΔ) διαφόρων ευρωπαϊκών χωρών. Για το σκοπό αυτό, το Κεφάλαιο ασχολείται με τον εντοπισμό, τη συλλογή και την ανάλυση μεταβλητών, οι οποίες, βραχυπρόθεσμες ή μακροπρόθεσμες, ενδέχεται να έχουν αντίκτυπο στη διαμόρφωση των ΔΣΔ. Συνδυαστικά χρησιμοποιήθηκαν οι τεχνικές των Beale et al. και ΑΚΣ, που εφαρμόστηκαν ώστε να ληφθεί το βέλτιστο σύνολο μεταβλητών για τη μοντελοποίηση των ΔΣΔ. Η ανάλυση επικεντρώνεται σε 20 ευρωπαϊκές χώρες για τις οποίες χρησιμοποιήθηκε ένα σύνολο 20 υποψήφιων επεξηγηματικών μεταβλητών για την περίοδο 2001–2015.

Αρκετές έρευνες σχετικά με τη χρήση του ΜΕΤ ως εποπτευόμενης τεχνικής μείωσης διάστασης έχουν αναπτυχθεί με την πάροδο των ετών στον τομέα της στατιστικής και χημειομετρίας για σκοπούς παλινδρόμησης. Ωστόσο, η ΜΕΤ μπορεί να είναι μια απαιτητική διαδικασία, ειδικά στην περίπτωση πολυμεταβλητής πολλαπλής παλινδρόμησης λόγω των χαρακτηριστικών και της πολυπλοκότητας των δεδομένων. Στο Κεφάλαιο 5 παρουσιάζεται η πρόταση για χρήση της μεθόδου ΜΕΤ ως τεχνική επιλογής μεταβλητών στη γραμμική παλινδρόμηση για φασματικά σύνολα δεδομένων υψηλών διαστάσεων. Πιο συγκεκριμένα, τεκμηριώνεται η πρόταση εκμετάλλευσης των συντελεστών παλινδρόμησης που υπολογίζει η ΜΕΤ για τον εντοπισμό και την εξάλειψη ασήμαντων επεξηγηματικών μεταβλητών από

την ανάλυση. Με αυτόν τον τρόπο, είναι εφικτή η απομάκρυνση των στατιστικά μη σημαντικών μεταβλητών και η επίτευξή της βέλτιστης μοντελοποίησης σε σύγκριση με τη κλασσική ΜΕΤ. Επιπροσθέτως, η προτεινόμενη τεχνική επιτυγχάνει μια απλούστερη δομή μοντέλου τόσο στην μονομεταβλητή όσο και στην πολυμεταβλητή περίπτωση.

Το Κεφάλαιο 6 προτείνει και διερευνά μια εύρωστη και εύκολα ερμηνεύσιμη μεθοδολογία, που ονομάζεται Elastic Information Criterion (EIC), ικανή να συλλαμβάνει την πολυσυγγραμμικότητα με μεγάλη ακρίβεια και αποτελεσματικότητα και έτσι να παρέχει μια την ορθότερη δυνατή αξιολόγηση του σετ δεδομένων. Η απόδοση του διερευνάται μέσω προσομοιωμένων και πραγματικών δεδομένων. Το EIC μπορεί να θεωρηθεί ως μια μη εποπτευόμενη γραμμική τεχνική επιλογής μεταβλητών.

Το Κεφάλαιο 7 επιχειρεί να εντοπίσει και να αναλύσει, μέσω τεχνικών πολυμεταβλητής ανάλυσης, μεταβλητές που συνδέονται (συσχετίζονται) με το Ακαθάριστο Εγχώριο Προϊόν (ΑΕΠ) και είτε βραχυπρόθεσμα είτε μακροπρόθεσμα επηρεάζουν την διαμόρφωση του.

Στόχος του Κεφαλαίου 8 είναι η πρόταση μιας καινοτόμου προσέγγισης για ευέλικτη και ακριβή μοντελοποίηση βαθμολόγησης πιστοληπτικής ικανότητας με τη χρήση όχι μόνο οικονομικών παραγόντων αλλά και χαρακτηριστικών πιστοληπτικής συμπεριφοράς. Επιπλέον, προτείνεται μια πολυδιάστατη αλγοριθμική διαδικασία μείωσης διάστασης προκειμένου να αναδειχθούν οι στατιστικά σημαντικές μεταβλητές και κατ' επέκταση να δημιουργηθεί ένα αξιόπιστο μοντέλο πρόβλεψης για τη βαθμολόγηση της πιστοληπτικής ικανότητας των εταιριών. Η προτεινόμενη νέα διαδικασία εφαρμόζεται στο ελληνικό σύστημα ξεχωριστά για «μικρές» και «μεγάλες» επιχειρήσεις.

Publications and Conferences

The results of Chapter 1 are presented in a chapter entitled *The Impact of Multicollinearity on Big Data Multivariate Analysis Modeling*, co-authored by Kimon Ntotsis and Alex Karagrigoriou, which has been published in Applied Modeling Techniques and Data Analysis 1, Computational Data Analysis Methods and Tools.

The results of Chapter 4 are presented in the articles entitled *On the Modeling of Pension Expenditures in Europe* and *On the Multivariate Modeling of Pension Benefits*, co-authored by Kimon Ntotsis, Marianna Papamichail, Peter Hatzopoulos and Alex Karagrigoriou, which have been published in Communications in Statistics: Case Studies, Data Analysis and Applications and The European Actuary, respectively. Versions of this works have been presented in various international conferences and workshops, i.e., (i) 2nd Conference of the Romanian Society of Probability and Statistics, (ii) 18th Conference of the Applied Stochastic Models and Data Analysis, (iii) Bernoulli – IMS One World Symposium 2020, and (iv) 1st LabSTADA Statistics and Probability e-Day.

The results of Chapter 5 are under review for possible publication as an article entitled *The Utilization of Partial Least Squares for Variable Selection on NIR Spectra Data*, co-authored by Beki Elisavet, Alex Karagrigoriou and Kimon Ntotsis. Versions of this work have been presented in international conferences and workshops, i.e., (i) 7th Stochastic Modelng Techniques and Data Analysis International Conference, and (ii) 3rd LabSTADA Statistics and Probability e-Day.

The work of Chapter 6 is reported in the article under the title *Interdependency Pattern Recognition in Econometrics: A Penalized Regularization Antidote*, co-authored by Kimon Ntotsis, Alex Karagrigoriou and Andreas Artemiou, which has been published in Econometrics. Versions of this article have been presented in various national and international conferences, workshops, and seminars i.e., (i) 2021 Joint Statistical Meetings, (ii) Bernoulli – IMS 10th World Congress in Probability and Statistics, (iii) 14th International Conference of the ERCIM WG on Computational and Methodological Statistics, (iv) The 19th Conference of the Applied Stochastic Models and Data Analysis, (v) 5th International Conference on Mathematical Techniques in Engineering Applications, (vi) 2021 Virtual International Workshop of G.S.I., (vii) Cardiff University, and (viii) 2nd Lab-STADA Statistics and Probability e-Day.

The results of Chapter 7 are presented in the article entitled *A Comparative Study of Multivariate Analysis Techniques for Highly Correlated Variable Identification and Management*, co-authored by Kimon Ntotsis, Emmanouil-Nektarios Kalligeris and Alex karagrigoriou, which has been published in International Journal of Mathematical, Engineering and Management Sciences. A preliminary version of this article has been presented at the 1st LabSTADA Statistics and Probability e-Day.

The results of Chapter 8 are presented in the article entitled *Multilevel Dimension Reduction for Credit Scoring Modelling and Prediction: Empirical Evidence for Greece*, co-authored by Panagiota Giannouli, Alex Karagrigoriou, Christos Kountzakis and Kimon Ntotsis, which has been published in Communications in Statistics: Case Studies, Data Analysis and Applications. Versions of this article have been presented at (i) 6th Stochastic Modeling Techniques and Data Analysis International Conference,

and (ii) 4th International Conference on Mathematical Techniques in Engineering Applications.

Acknowledgements

After nearly four years of effort, today is the day: writing this note of thanks puts the finishing touches on my Doctoral Dissertation. It has been an intense period of learning for me, both scientifically and personally. This Ph.D. has had a substantial impact on me. I'd like to thank every individual who has supported and aided me during this challenging and stressful period.

First and foremost, I want to thank my principal supervisor, Professor Alex Karagrigoriou, for the guidance and encouragement he has provided me with a perfect blend of insight and humour. I'm proud of and grateful for the opportunity to work alongside a mentor. You always unintentionally motivate those around you to do better.

I am grateful to my co-supervisor, Professor Andreas Artemiou, for his ongoing advice and for his never-ending supply of fascinating projects in my research field over the past few years. His perspective on research, science, and life is inspirational.

Some special words of gratitude go to my friends, for always being a great source of encouragement when things got tough or I was feeling down. Thank you, Christina, Manolis, Thanasis, Thanos, Vasiliki, Vaso, and many others for sharing times of great anxiety as well as great excitement.

In addition I would like to thank all my professors, colleagues and collaborators for their contributions in my research throughout this period: Mrs Beki Elisavet (Univ. of the Aegean), Asst. Prof. Gaki Eleni (Univ. of the Aegean), Dr. Giannoulli Panagiota (Univ. of the Aegean), Assoc. Prof. Hatzopoulos Petros (Univ. of the Aegean), Dr. Kalligeris Emmanouil-Nektarios (Univ. of Rouen Normandy/Univ. of the Aegean), Asst. Prof. Kountzakis Christos (Univ. of the Aegean), Mrs Papamichail Marianna (National Actuarial Authority/Univ. of the Aegean), Asst. Prof. Parpoula Christina (Panteion Univ. of Social and Political Sciences), Asst. Prof. Rakitzis Athanasios (Univ. of Piraeus), and Assoc. Prof. Tsilika Kyriaki (Univ. of Thessaly).

Last but definitely not least, I'd like to thank my parents, Ioannis and Anna, as well as my sister Angeliki, for their wise counsel and sympathetic ear. You are always there for me, and I could not imagine being here and doing what I love if you had not made such (statistical) significant sacrifices.

Contents

List of Figures

List of Tables

List of Abbreviations

2D	**Two-D**imensional
3D	**Three-D**imensional
AIC	**A**kaike **I**nformation **C**riterion
ANOVA	**AN**alysis **O**f **VA**riance
BD	**B**ig **D**ata
BDA	**B**ig **D**ata **A**nalytics
BHHJ	**B**asu, **H**arris, **H**jort, and **J**ones divergence measure
BIC	**B**ayesian **I**nformation **C**riterion
BoT	**B**alance **o**f **T**rade to GDP
CAB	**C**urrent **A**ccount **B**alance
CCA	**C**anonical **C**orrelations **A**nalysis
CCS	**C**redit **C**onsolidation **S**ystem
CFS	**C**orrelation-based **F**eature **S**election
CI	**C**onditional **I**ndex
CPI	**C**onsumer **P**rice **I**ndex
CSM	**C**redit **S**coring **M**odelL(ing)
DA	**D**ata **A**ugmentation
DF	**D**egrees of **F**reedom
DFO	**D**efault **F**inancial **O**bligation System
DIC	**D**ivergence **I**nformation **C**riterion
DR	**D**imensionality **R**eduction
DRT	**D**imensionality **R**eduction **T**echniques
EIC	**E**lastic **I**nformation **C**riterion
EN	**E**lastic **N**et
ENR	**E**lastic **N**et **R**egularization
Eurostat	**Stat**istical Office of the **Euro**pean Communities
EU	**E**uropean **U**nion
FS-PLS	**F**eature **S**election **P**artial **L**east **S**quares
GDP	**G**ross **D**omestic **P**roduct
GDP$_{GR}$	**G**ross **D**omestic **P**roduct **G**rowth **R**ate
GLM	**G**eneralized **L**inear **M**odels
GovDebt	**Gov**ernment **Debt** to GDP
ICA	**I**ndependent **C**omponent **A**nalysis
Inf$_R$	**Inf**lation **R**ate
Int$_R$	**Int**erest **R**ate
KL	**K**ullback-**L**eibler
KMO	**K**aiser-**M**eyer-**O**lkin measure of sampling adequacy
Lasso	**L**east **A**bsolute **S**hrinkage and **S**election **O**perator
LDA	**L**inear **D**iscriminant **A**nalysis
LV	**L**atent **V**ariables
MDIC	**M**odified **D**ivergence **I**nformation **C**riterion
MDS	**M**ulti**D**imensional **S**caling

MLE	**M**aximum **L**ikelihood **E**stimation
MLR	**M**ultiple **L**inear **R**egression
M-PLSR	**M**ultivariate **P**artial **L**east **S**quares **R**egression
MPS	**M**ortgages and **P**re-notations to Mortgages **S**ystem
MSE	**M**ean **S**quared **E**rror
M-PLS	**M**ultivariate **P**artial **L**east **S**quares
M-PLS	**M**ultivariate **P**artial **L**east **S**quares
M-PLSR	**M**ultivariate **P**artial **L**east **S**quares **R**egression
NRV	**N**eighbor **R**etrieval **V**isualizer
NIPALS	**N**onlinear **I**terative **P**artial **L**east **S**quares
NIR	**N**ear **I**nf**R**ared
OECD	**O**rganisation for **E**conomic **C**o-operation and **D**evelopment
OLS	**O**rdinary **L**east **S**quares
O-PLS	**O**rthogonal **P**rojections to **L**atent **S**tructures
PAYG	**P**ay **A**s **Y**ou **G**o
PCA	**P**rincipal **C**omponent **A**nalysis
PCR	**P**rincipal **C**omponent **R**egression
PCs	**P**rincipal **C**omponents
PLS	(univariate) **P**artial **L**east **S**quares
PLSR	**P**artial **L**east **S**quares **R**egression
PPE	**P**ublic **P**ension **E**xpenditures as percentage of GDP
PRESS	**PR**edicted **E**rror **S**um of **S**quares
RMSECV	**R**oot **M**ean **S**quare **E**rror of **C**ross **V**alidation
SDR	**S**ufficient **D**imensionality **R**eduction
S.E./s.e.	**S**tandard **E**rror
SFA	**S**low **F**eature **A**nalysis
SIMPLS	**S**tatistically **I**nspired **M**odification of the **P**artial **L**east **S**quares
SME	**S**mall to **M**edium-sized **E**nterprises
step	**Step**wise regression
stepAIC	**Step**wise regression based on **AIC**
stepBIC	**Step**wise regression based on **BIC**
stepDIC	**Step**wise regression based on **DIC**
stepR^2_{adj}	**Step**wise regression based on **step**R^2_{adj}
SVD	**S**ingular **V**alue **D**ecomposition
TOL	**T**olerance **L**imit
Unem$_R$	**Unem**ployment **R**ate
VAT	**V**alue-**A**dded **T**ax
VIF	**V**ariance **I**nflation **F**actor
WoE	**W**eight **o**f **E**vidence
WR	**W**old's **R**atio

Chapter 1

Introduction to Dimensionality Reduction

1.1 The issue of multicollinearity

The tremendous increase in the development of technology as well as the creation of new databases on a variety of topics makes Big Data Analytics (BDA; BD for Big Data) more efficient to work with. However, more is not always better. Large amounts of data might sometimes fail to perform properly in data analytics applications. Indeed, when it comes to modelling, a multitude of explanatory variables for extensive time periods can cause inconsistencies in the interpretation of statistical results. The most important obstacle that one has to overcome is the existence of multicollinearity between the variables.

Nowadays, researchers involved in statistical modelling processes, are able to collect data on multiple factors over a long time horizon, which may contain millions of observations. Trying to model a variable having such a large dataset, can cause additional problems (such as inaccurate and disorderly databases, computational complexity, and insufficient analytical skills) as opposed to modelling with smaller datasets. Due to such problems, there was the need for the creation of a particular category of data analysis, named BDA, which is a very growing branch of data science. As IBM states, "*Big data analysis is the use of advanced analytic techniques against very large, diverse data sets that include structured, semi-structured, and unstructured data of different sizes. Big data is a term applied to data sets whose size or type is beyond the ability of traditional relational databases to capture, manage, and process the data with low latency. And it has one or more of the following characteristics – high volume, high velocity, or high variety*" IBM (2021).

The most common consequence of such a big dataset modelling in statistical analysis is the existence of multicollinearity, which is defined as the high linear association between two or more variables (i.e., one or more variables can be linearly predicted from other(s) at a notable degree of accuracy) (Alin (2010)). Not so often it can be caused by the false use of dummy variables in the modelling, the repetition of the same kind of variable, or when the data are insufficient, which can be resolved by collecting more data. The coexistence of these cohort variables in a regression analysis can result in inconclusive or even incorrect interpretation, while it may affect the forecasting process (Bayer (2018)). Even though relatively small multicollinearity may cause no harm, moderate and severe one can abate the statistical power of the regression and lead to overfitting due to variables redundancy. That phenomenon is quite common nowadays due to the size of unfiltered information. This results in datasets with variables that are, to a significant extent, collinear due

to interrelationships that lurk and potentially lead to misleading models (overestimation/underestimation) and inaccuracy in parameter estimation.

In the presence of multicollinearity, the proper interpretation of the data analysis may not be reliable (Silvey (1969)). Due to this fact, multicollinearity has reached the point of being an additional "fifth" assumption in the case of multiple and multivariate regression, among the already existing ones - normality, homoscedasticity, and independence of the residuals and linearity between the dependent(s) and each independent variable (Yoo et al. (2014)).

Multicollinearity falls into one of the following two categories namely, structural and data-based multicollinearity. Structural one, also known as perfect multicollinearity, occurs when a byproduct variable exists in the dataset along with the one that originates from and can be expressed (assuming that correlation exist between at least two variables of the total p in a given X-matrix) as:

$$\sum_{j=1}^{p} a_j X_j = 0$$

where, there are $a_j \neq 0$ that confirms the equation. In other words, is a mathematical artifact caused by generating predictors with the use of already existing ones. Due to the fact that a perfect collinear relationship between the variables included in the model exists, $X^\intercal X$ becomes singular and thus it is not feasible to use the Ordinary Least Squares (OLS) regression to estimate the value of the parameters due to $X^\intercal X$ non-invertibility . Therefore, perfect multicollinearity violates one of the linear regression model assumptions.

Data-based, also known as high (also abbreviated as extreme or severe) multicollinearity, occurs between the variables in the original unprocessed dataset and is the most common type when it comes to observational experiments. In this case, the relation between the variables is approximately linear and can be written as:

$$\sum_{j=1}^{p} a_j X_j \approx 0$$

where there are $a_j \neq 0$ that confirms the equation. High multicollinearity often exists in big multivariate complex datasets, where variables may be quantified in dissimilar sized measures which can enhance the significance of insignificant variables and potentially conceal the statistically significant ones (Ueki and Kawasaki (2013), Yue et al. (2019)). In this case, where $(X^\intercal X)^{-1}$ computation is possible, the following situations can arise in a modelling process: (i) It becomes difficult to distinguish the effects of predictors on the response variable during a statistical modelling process. When the rest of the predictors remain unchanged, a regression coefficient expresses the impact of a one-unit change in the corresponding predictor. When predictors are correlated, they share a common variability pattern, which means they both increase and decrease at the same time. As a result, estimating how much-correlated predictors affect the response variable is difficult and therefore, the apparent effects are deceptive. (ii) Furthermore, when multicollinearity is presented, there are numerous combinations of estimated coefficients that all produce similar predictions. This means that the estimates have a high standard error. As a result, the T-statistic $t = \hat{\beta}_j / s.e.(\hat{\beta}_j)$ decreases, lowering the power of the hypothesis test $H_0 : \beta_j = 0$. This means that zero coefficients may not be detected as such but may appear statistically significant, resulting in the existence of redundant variables. In these cases, the coefficient of determination -the measure used to evaluate a model's goodness

of fit (i.e., performance)- will be high, and the produced model will be interpreted as adequate. However, redundancy causes overfitting of the regression model, which results in the regression model failing to perform adequately for samples. (iii) Additionally, because the produced coefficient estimates are imprecise, predictions of samples located outside the space covered by the training set will be imprecise as well.

However, multicollinearity is not always harmful. There are cases that can be categorized as non-existing, small, and moderate (medium). In these cases, there is no need for any process for its elimination. Usually, insufficient data may guide the deceitful existence of multicollinearity (Ntotsis and Karagrigoriou (2021)).

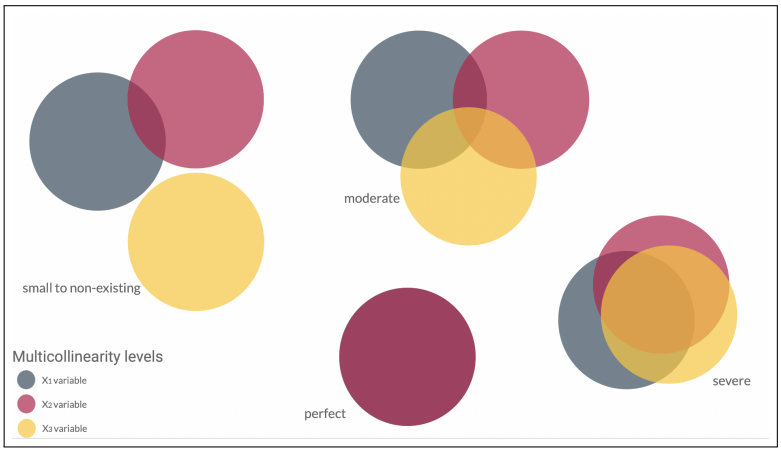


FIGURE 1.1: Triple Venn diagram demonstrating the variation of multicollinearity

Figure 1.1 displays the five states of multicollinearity. From small multicollinearity between variables X_1 and X_2 while no existent one between X_3 and each of them in the upper left Venn diagram, to moderate multicollinearity between all three in the upper right corner, to severe and perfect multicollinearity in figures' lower half.

Perfect multicollinearity is highly uncommon and the easiest to handle and avoid by a thorough examination of the model's variables. However, high multicollinearity -the most habitual, can cause severe estimation and interpretation problems. The most prevalent consequence which appears in the presence of multicollinearity is overfitting in regression analysis modelling due to the redundancy of variables, which reduces the power of the model to identify the statistically significant variables. That means that the model is too complex and the model's measures of assessment, such as the coefficient of determination, are misleading because instead of describing the proportion of the variance in the dependent variable that is predictable from the independent variables, describe the random error in the data. It is also possible that parameter estimates may not accurately describe the impact of the associated variables on the dependent variable. It can also result in the alteration of the indicator and the immensity of the partial regression coefficients from one sample to another. Furthermore, although that phenomenon seems quite paradoxical, reports have been made for non-compatible results between F-test and T-test when multicollinearity exists (Largey (1996)).

1.1.1 Review of multicollinearity measures

To detect the multicollinear variables in a set of data and eliminate them, assorted criteria have been developed over time. Some of these are briefly presented below. An easy way to have a first glimpse at multicollinearity's state is through a correlation matrix. Relatively high correlations indicate the possible existence of multicollinearity. Dissimilar results about the coefficients between F and T-tests (i.e., in the regression model, the F-test comes to the conclusion that at least one variable is statistically significant, while at the same time the T-test suggests that none of the variables actually is) is an indicator for the existence of harmful multicollinearity. Although that phenomenon seems quite paradoxical, it has been thoroughly examined and explained by Largey (1996), as they provide two reasons for the occurrence. The first one is the existence of multicollinearity in the model, in which the existence of a relationship can be established but not the individual influence of each variable. The second reason stems from the value of the degrees of freedom (DF) of the residuals. If the residuals DF are $\geqslant 3$, then, the significant point of $F_{(k,\, n-k-1)}$ is lower than the significant point of $F_{(1,\, n-k-1)}$ which corresponds to the significant point of the t-statistics. Hence, when all t-statistics are equal or approximately so, they may all be non-significant while F is significant. The explanation is that a significant F-ratio does not indicate the significance of any given regression coefficient but merely the existence of at least one linear combination which is significantly different from zero. Additionally, significant R-squared shifts when variables are inserted/removed can also imply the existence of severe multicollinearity (Geary and Leser (1968)).

There are several partially robust criteria and indices for multicollinearity detection focusing either on the coefficient of determination and similar measures or on the eigensystem analysis. Some of the most regularly used are:

Collinearity diagnostics such as eigensystem analysis and Conditional Index (CI) (Belsley (1991)) can highlight the issue of multicollinearity. Correlation matrix-based eigenvalues near zero presuppose multicollinearity among the variables (Hair et al. (2010), Kendall (1957)), while if the CI of Equation 1.1 is greater than 10, empirically, one can say that it leads to the same conclusion (Belsley (1991), Hair et al. (2010)).

$$CI_j = \sqrt{\frac{\lambda_{max}^{ev}}{\lambda_j^{ev}}}, \tag{1.1}$$

where λ_j^{ev} is the eigenvalue emerged from the original variables correlation matrix, λ_{max}^{ev} is the maximum eigenvalue, $j = 1, 2, \ldots, k$ is the number of variables and $\lambda_1^{ev} \geq \lambda_2^{ev} \geq \cdots \geq \lambda_k^{ev}$.

Besides, Kovács et al. (2005) used eigensystem analysis to compose the Red indicator, presented in Equation 1.2, for proper detection. When the indicator approaches zero, then multicollinearity is low, while when it approaches 1, then it can be considered high.

$$Red = \frac{\frac{\sqrt{\sum_{j=1}^{k}(\lambda_j^{ev}-1)^2}}{k}}{\sqrt{k-1}}. \tag{1.2}$$

Farrar-Glauber test (Farrar and Glauber (1967)) approaches the issue with the comprised of a 3-test procedure that examines the presence of multicollinearity, the existence of collinear regressors, and the form of their affiliation. They also proposed the use of a measure based on the ratio of explained to unexplained variance (Farrar and Glauber (1967)), the large values of which indicate multicollinearity.

$$w_j = (r^{jj} - 1) \times \left(\frac{n-k}{k-1} \right), \tag{1.3}$$

where $r^{jj} = \frac{1}{1-R_j^2}$ and R_j^2 is the R-squared of the auxiliary regression of each j variable against all the others.

Klein (1962) and Theil (1971) independently proposed rules based on R_j^2, and its impact on the overall R-squared. Klein states that if R_j^2 surpasses the overall R^2, then multicollinearity can be worrisome. On the other hand, Theil's rule asserts that if the resulting m from Equation 1.4 is 0, then multicollinearity is absent, while if it is approximately equal to 1, then it can be considered troublesome.

$$m = R^2 - \sum_{j=1}^{k} (R^2 - R_{-j}^2), \tag{1.4}$$

where R_{-j}^2 is the resulting R^2 of the full model without the inclusion of the X_j variable.

Leamer (Greene (2002)) suggested a method based on the variance of the estimated coefficients:

$$C_j = \left(\frac{\left(\sum_{j=1}^{k} (X_{ij} - \bar{X}_j)^2 \right)^{-1}}{(X'X)_{jj}^{-1}} \right)^{\frac{1}{2}}. \tag{1.5}$$

Equation 1.5 is used for ruling and takes values in [0,1]. When C_j approaches the left end, then multicollinearity exists; while, when it approaches the right end, then it can be considered non-existent. Although all the above are well established and frequently used techniques for multicollinearity detection, the criterion that is the most frequently used in various fields is the Variance Inflation Factor (VIF) (Gujarati and Porter (2008)) which uses the coefficient of determination for detection purposes and is formulated as follows:

$$VIF_j = \frac{1}{1 - R_j^2} = \frac{1}{\text{TOL}} \tag{1.6}$$

VIF indicates how magnified is the variance of an estimator in the presence of multicollinearity. When no multicollinearity among variables exists, then VIF_j = 1 and when R_j^2 approaches 1, then VIF_j approaches infinity. If VIF_j is greater than 5, then the j^{th} variable is considered multicollinear and is proposed for extraction for a better result interpretation (Gujarati and Porter (2008)). However, the acceptance range is subject to requirements and constraints, with most suggesting the acceptance threshold to be equal to 5 or 10. Disregarding its regular usage, VIF lags behind in some cases. More specifically, as Gujarati and Porter state (2008) *"high VIF is neither necessary nor sufficient to get high variances and high standard errors. Therefore, high multicollinearity, as measured by a high VIF, may not necessarily cause high standard errors"*. Tolerance Limit (TOL) is also a detection measure, closely related to VIF as it is its denominator. Weisburd and Britt (2013) state that a value under 0.2 indicates severe multicollinearity.

Lastly, the IND1 indicator proposed by Ullah et al. (2016), can be used for detection purposes. Its corresponding formula is

$$IND1_j = (R_j^2 - 1) \times \left(\frac{1-k}{n-k} \right), \tag{1.7}$$

and when $IND1_j \leq 0.02$, then multicollinearity exists. For more about multicollinearity measures, one can refer to Halkos and Tsilika (2018), and Imdadullah et al. (2016).

When multicollinearity is harmful, then all the aforementioned measures usually fail to recognize patterns among variables. This occurs as a consequence of model overfitting. There are several ways to deal with this issue. The most common solution when such states of multicollinearity exist is to remove the byproduct variables. However, these methods can be extremely time-consuming and not so trustworthy. Researchers nowadays tend to prefer the use of dimension reduction techniques that focus either on variable selection or extraction, either on coefficient penalization. Techniques such as Principal Component Analysis and Partial Least Squares are considered optimal for purpose fulfilment. Alternative approaches such as Least Absolute Shrinkage and Selection Operator, Ridge, Elastic Net Regularizations, etc., have been developed and utilized to address the issue of coefficient penalization absence and expansion multicollinearity. All these techniques will be thoroughly discussed in the following Chapters.

1.2 Dimensionality Reduction: The cure to the curse

The curse of dimensionality refers to phenomena that emerge when investigating and analysing data in high-dimensional spaces that do not occur in low-dimensional settings. The curse exists because as dimensionality increases, the sample needed decreases exponentially. The continuous adding of information without increasing the number of training samples will lead to the dimensionality of the feature space expanding and eventually to data becoming sparse. In order for the results in a high-dimensional dataset to be reliable, the dimensionality must grow exponentially with the available data, which is a rare phenomenon due to data dissimilarities. To address this curse, the multicollinearity, and any other potential side issues associated with BDA, special techniques, namely Dimensionality Reduction Techniques (also referred to as Dimension Reduction Techniques, DRT; DR for Dimension Reduction) were developed. The DRT are defined as techniques that converse and project a high-dimensional space dataset in a low-dimensional space while maintaining unaltered the variability (information) and properties of the data. Ideally, the proper utilization of DRTs can lead to the intrinsic dimension of a given dataset; i.e., the thin line between the maxima variable loss and the minima information lost, that can acknowledge the observable properties structure and representation of the input dataset. Thusly, it is feasible to have a more effective perception of the data by "*sacrificing*" a small portion of its original information. Figure 1.2 gives a brief visual aspect of how effective the DRT can be in the research area. The left graph illustrates the original data, while the right graph presents the results researchers retrieve via DR. In A can be seen, that the experiments have been grouped via the neighbor retrieval visualizer DRT. This technique also utilizes colors and widths in the component boxes to embed the complex information that exists in the original data. A detailed interpretation of the experiment, one can find in Honeine et al. (2018).

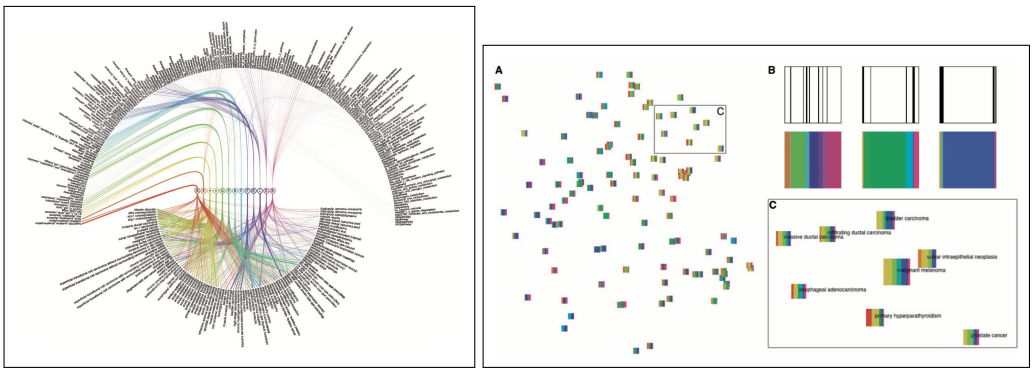


FIGURE 1.2: The use of a circular link diagram for a visual information retrieval interface to a collection of microarray experiments visualized as glyphs on a plane.
Left Figure: Original data. Right Figure: A: Glyph locations have been optimized by the neighbour retrieval visualizer so that relevant experiments are close by. For this experiment data, relevance is defined by the same data-driven biological processes being active, as modelled by a latent variable model (component model). B: Enlarged view with annotations; each colour bar corresponds to a biological component or process, and the width tells the activity of the component. These experiments were retrieved as relevant for the melanoma experiment shown in the centre. C: The biological components (nodes in the middle) link the experiments (left) to sets of genes (right) activated in them.
Source: Honeine et al. (2018)

Due to the large amount of data existing nowadays, dimension reduction has become a must-have tool for researchers in order to analyse their data. The benefits that DRT accomplishes prevail over the information loss that requires. Those beneficial characteristics of DRT can be summarized in eight aspects as presented below:

- Overfitting is avoided by reducing dimensionality. When there are many features in the data, the models become more complex and tend to overfit the training data.
- Data visualization. When we reduce the dimensionality of higher dimensional data into two or three components, the data can easily be plotted on a two-dimensional (2D) or three-dimensional (3D) plot.
- Multicollinearity is addressed by dimensionality reduction. In regression, multicollinearity occurs when an independent variable is highly correlated with one or more of the other independent variables. Dimensionality reduction exploits this by combining highly correlated variables into a set of uncorrelated variables.
- A lower number of dimensions in data means less training time and computational resources, which improves the overall performance of machine learning algorithms -Machine learning problems with many features require extremely slow training. The majority of data points in high-dimensional space are very close to the space's boundary. This is due to a large amount of space available in high dimensions. Most data points in a high-dimensional dataset are likely to be far apart. As a result, algorithms are unable to train effectively and efficiently on high-dimensional data.

- DR is highly suitable for factor analysis. It can be an effective technique for identifying latent variables that are not directly measured in a single variable but rather inferred from other variables in the dataset. These latent variables are referred to as factors.
- Dimensionality reduction reduces data noise. DR reduces data noise by keeping only the most important features and removing redundant features. This will improve the accuracy of the model.
- DR can be used for image compression. Image compression is a technique that reduces the size of an image in bytes while retaining as much of the image's quality as possible. The pixels that comprise the image can be thought of as image data dimensions (columns/variables).
- DR can be used to convert non-linear data into linearly separable data.

However, even though the beneficial characteristics of DR can lead to robust and efficient analysis results, they might be disadvantageous too. Improper selection or implementation of a DRT can potentially lead to significant information loss. The majority of statistical theories and applications dealing with DR are focused on linear DR, which in many cases is undesirable or non-existing. Finally, when it comes to feature extraction, most techniques are a bit "abstract", meaning they apply a rule of thumb in order to obtain the new variables. That brings up the question, of which new variables to retain.

1.2.1 Approaches to DRT

The DR techniques can be classified/categorized based on several criteria. Most prevailed differentiations between the approaches are:

Linear and non-linear DRT

This distinction is based on data characteristics. When linear relations between variables occur, then the data is transformed to a low dimension space as a linear combination of the original variables using linear dimensionality reduction. When the data is in a linear subspace, the original variables are replaced by a smaller set of underlying variables. Such approaches include principal component analysis (PCA), multidimensional scaling (MDS), partial least squares (PLS), linear discriminant analysis (LDA), canonical correlations analysis (CCA), independent component analysis (ICA), slow feature analysis (SFA), singular value decomposition (SVD), Neighbor Retrieval Visualizer (NRV), and sufficient dimensionality reduction (SDR). When the original high-dimensional data contains non-linear relationships, nonlinear dimensionality reduction is used. The data is represented in a lower dimension while the original distances between the data points are preserved. Such approaches include kernel principal component analysis, Sammon's mapping, local linear embedding, isomap, Laplacian eigen map, and uniform manifold approximation and projection are among the most regularly used.

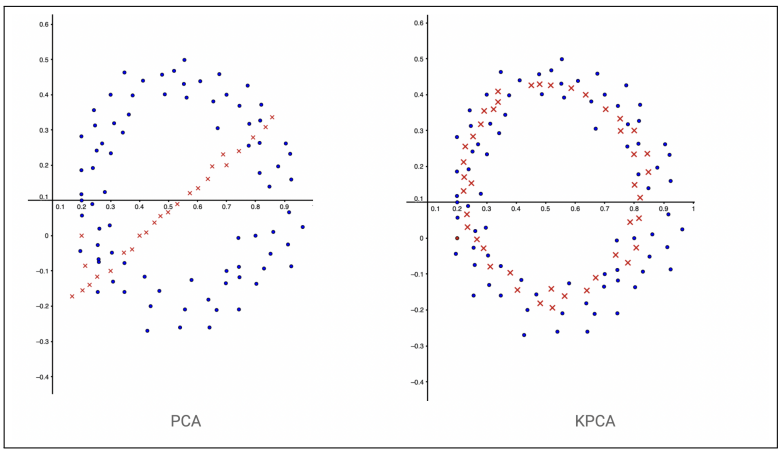


FIGURE 1.3: Linear (PCA) vs non-linear (KPCA) DRT

According to the comparison in the Figure 1.3, KPCA produces an eigenvector with higher variance (eigenvalue) than PCA. Because KPCA is a circle and PCA is a straight line for the largest difference of the projections of the points onto the eigenvector (new coordinates), KPCA has a higher variance than PCA.

Feature extraction and feature selection DRT

Feature (variable) selection methods include algorithms that aim to find irrelevant and/or redundant variables of a dataset. Then, these variables are removed. A new dataset has a lower dimension, as it consists of a subset of the variables of the initial dataset (Guyon and Elisseeff (2003)). In most techniques in this category, all variables are matched with a value, arising from a criterion. According to its value, every variable is evaluated and it is decided either on its removal or its selection. The physical meaning of retained variables does not change at all. Despite the advantage of interpretability, information captured in interactions and correlations between selected and removed variables is lost (Li and Zeng (2009)). Some popular techniques of this kind are Information Gain, Relief, Fisher Score, Forward Feature Selection, Chi-square Test, Backward Feature Elimination, Least Absolute Shrinkage and Selection Operator (Lasso), and Elastic Net (EN). On the contrary, the application of feature extraction (projection) techniques results in the transformation of a dataset -data points are projected to low dimensional space. This is achieved with the use of original variables as elements of combinations that summarize information from initial variables (Li and Zeng (2009)). This results in the newly generated variables, (also mentioned as components or latent variables in the literature) being correlated with the original ones. Among the most applied techniques are PCA, PLS, CCA, and LDA.

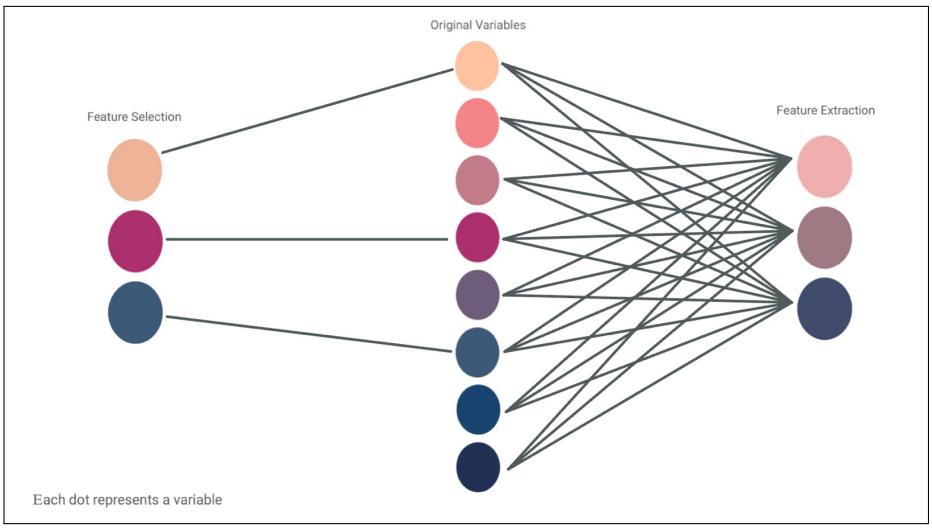


FIGURE 1.4: The impact of Feature Selection and Feature Extraction on the original variables

As can be seen in Figure 1.4, when in feature selection, an amount of the original variables are being selected, while when in feature extraction all variables are utilized for the generation of each of the components with different weights. As can be seen in the graph, the components have different colours; this is due to the fact that even though each one contains all original variables, some play a more significant role in the formation of the component than others.

Supervised and unsupervised DRT

The majority of DRT are unsupervised learning techniques. The distinction between the above categories lies in the existence of supervised information (class labels) or not. LDA and PLS are considered supervised techniques since the first extracts the optimal discriminant vectors when labels are available and the latter uses the response variable(s) in order to obtain the latent variables of the model. The most recognized unsupervised techniques are considered to be PCA and kernel PCA, which try to maintain the data structure without acknowledging the existence of labels. Several other subcategories of supervision also exist, such as semi-supervised DRT -which learns from both labelled and unlabelled data, linear and non-linear (un)supervised DRT, etc.

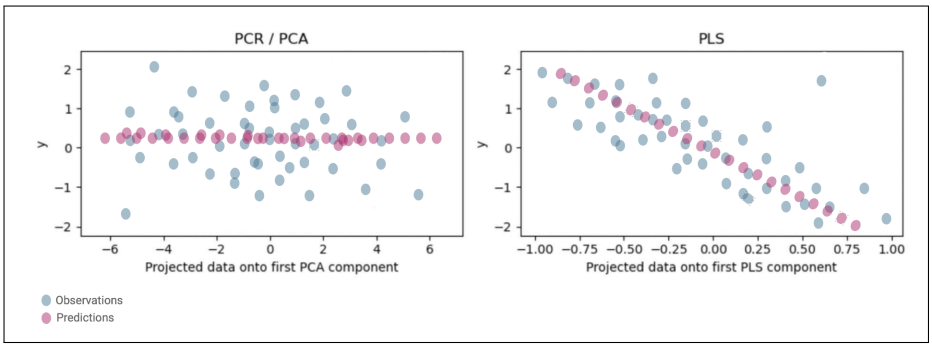


FIGURE 1.5: Unsupervised (PCA) and Supervised (PLS) DRT

Figure 1.5 compares PCA and PLS regression based only on the first component. The plot displays the projected data onto the first component against the observations (original data points) in both models. In both cases, the regressors will use this projected data as training data. Despite being the most predictive direction, the unsupervised PCA transformation dropped the second component, i.e. the direction with the lowest variance. Because PCA is an entirely unsupervised transformation, projected data have low predictive power on the target. The PLS regressor, on the other hand, captures the effect of the direction with the lowest variance because it uses target information during the transformation: it recognizes that this direction is actually the most predictive. We notice that the first PLS component is negatively correlated with the target, which is due to the arbitrary nature of eigenvector signs. A similar example can be found in Pedregosa et al. (2011).

In this Thesis, the unsupervised DR technique PCA is being thoroughly investigated in various data types on its own and in combination with other techniques; almost exclusively for modelling purposes.

Chapter 2 documents in detail the theoretical framework behind PCA (Section 2.1) and its supervised equivalent, PLS method (Section 2.2). Both techniques are considered optimal for feature extraction due to their ability to obtain DR when high-dimensional datasets are being analysed. A theoretical comparison between those two takes place, as well as a comparison concerning their benefits and disadvantages (Section 2.3). Additionally, a more contemporary approach to DR, through regularization techniques is also documented in the final part of (Section 2.4).

Chapter 3 mentions criteria and indices implemented in Chapter 4 – Chapter 8 for assessing the power of models resulted through the analyses considered.

The purpose of Chapter 4 concerns the modelling of public pension expenditures (PPE) of various European countries. For this purpose, the Chapter deals with the identification, collection, and analysis of variables, which, either short-term or long-term, may have an impact on the shaping of PPE. A mixture of the Beale et al. (1967) technique and PCA was implemented to obtain the optimal set of variables for the modelling of PPE. The analysis focuses on 20 European countries for which a set of 20 possible explanatory variables for the period 2001–2015, were used.

Several works concerning the utilization of PLS as a supervised dimension reduction technique have been developed over the years in the field of chemometrics, among others, for regression purposes. However, PLS can be a challenging procedure, especially in the case of multivariate multiple regression due to data characteristics and complexity. Thus, in Chapter 5 the proposal of Feature Selection PLS (FS-PLS) takes place. FS-PLS is a PLS-based method that acts as feature selection and feature extraction technique at the same time and is utilized in linear regression tasks that involve high dimensional spectral data sets. More precisely, the suggestion to exploit the regression coefficients that PLS estimates in order to identify and eject the insignificant predictor variables from the analysis is documented. In such a manner, we are able to remove the uninformative variables and obtain, in most cases, better results than classical PLS regression but with a simpler structure. The comparison of the proposed algorithm with the classical PLS and PCA occurs in both univariate and multivariate regression scenarios.

When it comes to variable interpretation, multicollinearity is among the biggest issues that must be surmounted, especially in this new era of BDA. Since even moderate size multicollinearity can prevent proper interpretation, special diagnostics must be recommended and implemented for identification purposes. Nonetheless, in the areas of econometrics and statistics, among other fields, these diagnostics are

controversial concerning their "successfulness". It has been remarked that they frequently fail to do proper model assessment due to information complexity, resulting in model misspecification (Lindner et al. (2020)). Chapter 6 proposes and investigates a robust and easily interpretable methodology, termed Elastic Information Criterion (EIC), capable of capturing multicollinearity rather accurately and effectively and thus providing a proper model assessment. Performance is investigated via simulated and real data. EIC can be considered an unsupervised linear feature selection technique.

Chapter 7 attempts to locate and analyse via multivariate analysis techniques, highly correlated variables which are interrelated with the Gross Domestic Product (GDP) and therefore are affecting either a short-term or a long-term shaping. For the analysis, three variable selection/extraction techniques were used. The case study focuses on annual data for Greece from the period 1980 to 2018.

The objective of Chapter 8 is the proposal of an innovative approach to flexible and accurate credit scoring modelling with the use of not only financial but also credit behavioural characteristics. In addition, we propose a multidimensional reduction algorithm in order to divulge the statistically significant variables that prevail and as an extension to create a reliable prediction model for credit scoring based on the effective combination of PCA and regularization methods. The proposed novel procedure is applied to the Greek system separately for small and large enterprises with the use of a Credit Bureau database with more than 200,000 cases.

Chapter 2

Review of Dimensionality Reduction Techniques

2.1 Principal Component Analysis

As mentioned prior, large amounts of data might sometimes fail to perform properly in data analytics applications and can cause inconsistencies in the interpretation of the results. In order to overcome this problem, which most likely comes as a result of the existence of multicollinearity, various DRT were developed. PCA is able to reduce the number of random variables, -under specific conditions and constraints, and create a new smaller set of variables based on the original one. Through this process, it is easier to interpret different statistical tests without losing the accuracy of the original variables in the sense that these techniques are intended to retain variation unchanged as much as possible.

The Beale et al. (1967) technique considers a more simplified version of PCA and is summarized by the following three-step procedure for discarding variables in multivariate analysis.

1 Locate the minimum eigenvalue and the corresponding eigenvector of the variance-covariance or correlation matrix.

2 Locate the element of the eigenvector with the highest absolute value. This value corresponds to a variable which will be removed from the model.

3 Repeat the above steps until *p-k* variables have been removed.

where p is the number of all variables and k is the number of eigenvalues that are greater than one. However, when a large amount of data is involved or a more complex multicollinearity structure exists, these techniques lead to model overestimation or underestimation.

PCA is a commonly applied DRT, introduced by Pearson (1901) and Hotelling ((1933), (1936)). It is a multivariate technique in which a data matrix X (X-matrix), which includes correlated variables, is transformed into a new one. Variables in the new matrix, also called Principal Components (PCs), are uncorrelated and ordered so as to contain the variance of the original X-matrix on a declining scale, starting from the first one. The beneficial property of the new matrix is that most variation of the X-matrix is compressed in the first few new variables. These variables, the number of which is selected by the user, form a low-dimension matrix, an approximation of X that can be used for modelling purposes.

As Varmuza and Filzmoser state (2009) *"PCA can be seen as a method to compute a new coordinate system formed by the latent variables, which is orthogonal, and where only the most informative dimensions are used."* From a geometrical point of view, an

X-matrix is projected/mapped to a new space (hyperplane, plane, or line), the coordinate system of which is formed by the PCs, which are oriented in the direction of maximized variance of data points. *"The coordinates of the samples in the new space are called scores, often indicated with the symbol T. The new dimensions are linear combinations of the original variables and are called loadings (symbol P)."* (Wehrens (2011)).

The PCA of an X-matrix of size $n \times p$ is:

$$X = T_p P_p^\intercal$$

where T is the matrix containing the scores of the samples, P is the matrix containing the loadings, and superscript $^\intercal$ indicates the transpose of a matrix. Subscript p indicates the number of latent variables that can be computed. However, as only a few Principal Components are almost always used for modelling since they suffice to explain most of the variance included in X, the original matrix can be written as:

$$X = T_m P_m^\intercal + E$$

where m, ranging from 1 to p, indicates the number of selected latent variables and E is the matrix containing the residual error. Geometrically, that is the perpendicular distance of each point onto the hyperplane formed by loading vectors. These quantities represent the loss of information because of the projection of X data points into a low-dimension space. Finally, the new, low-dimension matrix can be written as:

$$\tilde{X} = T_m P_m^\intercal$$

where $\tilde{X}$ indicates the approximation of X, that can be used for modelling purposes discharged of noise.

2.1.1 Steps to build a PCA model

The first step in PCA is centring data matrix X in order to remove arbitrary bias from measurements. This is achieved by replacing each x_{ij} element by:

$$x_{ij} - \bar{x}_j$$

where $(\bar{x}_j)$ indicates the mean value of column j. After this process, in the mean-centred matrix X, each column has a mean of zero. This technically means that data points have been moved to the centre of the coordinate system while the distances between them do not change at all.

In some cases, datasets include variables of different magnitudes, because they are measured in different units. As a result, some variables have different statistical weights in the analysis. This problem can be solved by replacing each x_{ij} element by:

$$\frac{x_{ij} - \bar{x}_j}{s_j}$$

where s_j indicates the standard deviation of the j_{th} variable, a process called scaling. In this way, the final columns in X have a mean of zero and a unit variance, and it should be noted that in this case the relative distance between data points is changed. However, if predictors are measured in the same units, scaling could cause the inflation of noise in uninformative variables (Wehrens (2011)).

The next step is to compute the matrices for Principal Components.

Singular Value Decomposition

A commonly applied technique to compute scores and loadings is the SVD, according to which the mean-centred X-matrix is decomposed as:

$$X = UDV^{\intercal}$$

where U is a matrix of size $n \times n$ and its columns are the left singular vectors of X, while V is a matrix of size $p \times p$ and its columns are the right singular vectors of X. Matrices U and V are orthogonal, meaning each column is orthogonal to the others. Matrix D is a diagonal $n \times p$ matrix, where diagonal elements d_i are related to variances of corresponding PCs. These quantities can be computed by:

$$\lambda_i = \frac{d_i^2}{n-1}$$

Finally, matrices U, D, V, T, and P are related to each other as follows:

$$X = (UD)V^{\intercal} = TP^{\intercal}$$

meaning that the matrix of loadings P is set equal to matrix V, while the matrix of scores T is set equal to matrix $[UD]$.

Eigen Decomposition

In the case of data sets with many original variables, the SVD process is considered computationally demanding and is avoided. Instead, another method called eigen decomposition is applied to either covariance matrix Σ or correlation matrix $\rho(X)$.

Each element in a covariance matrix represents the covariance between two variables, a quantity that measures their joint variability of them, and it is computed by:

$$cov(X_1, X_2) = E[(X_1 - E(X_1))(X_2 - E(X_2))]. \tag{2.1}$$

Given a data set that includes $X_1, X_2, \ldots, X_p$ variables, the covariance matrix is symmetric as shown below:

$$\Sigma = \begin{bmatrix} cov(X_1, X_1) & cov(X_1, X_2) & \ldots & cov(X_1, X_p) \\ cov(X_2, X_1) & cov(X_2, X_2) & \ldots & cov(X_2, X_p) \\ \vdots & \vdots & & \vdots \\ cov(X_p, X_1) & cov(X_p, X_2) & \ldots & cov(X_p, X_p) \end{bmatrix}$$

Diagonal elements of the Σ matrix represent variances of variables, since by definition

$$Var[X_1] = E[(X_1 - E(X_1))^2] = E[(X_1 - E(X_1))(X_1 - E(X_1))] = cov(X_1, X_1).$$

As a result, this matrix is also called the variance-covariance matrix.

However, in PCA, the original X variables are mean-centred and Equation 2.1 becomes:

$$cov(X_1, X_2) = E[X_1 X_2].$$

In terms of a matrix, that includes all p variables, this can be written as $E[X^\intercal X]$ and when the variance-covariance matrix refers to a sample data set, it is equal to the $X^\intercal X$ matrix.

Correspondingly, each element in the correlation matrix represents a correlation between two variables, a measure of the linear relationship between them, and it is computed by:

$$cor(X_1, X_2) = \frac{cov(X_1, X_2)}{\sqrt{Var[X_1]}\sqrt{Var[X_2]}}.$$

Given a data set that includes $X_1, X_2, \ldots, X_p$ variables, the correlation matrix is symmetric and diagonal elements are equal to one, as shown below:

$$\rho(X) = \begin{bmatrix} 1 & cor(X_1, X_2) & \ldots & cor(X_1, X_p) \\ cor(X_2, X_1) & 1 & \ldots & cor(X_2, X_p) \\ \vdots & \vdots & & \vdots \\ cor(X_p, X_1) & cor(X_p, X_2) & \ldots & 1 \end{bmatrix}$$

since

$$cor(X_i, X_i) = \frac{cov(X_i, X_i)}{\sqrt{Var[X_i]}\sqrt{Var[X_i]}} = \frac{Var[X_i]}{Var[X_i]} = 1, \forall i.$$

It should be noted that correlation is independent of the scales of variables, while covariance is not. This is the reason why the correlation matrix is used when variables have different measurements; oppositely, the covariance matrix is used when all variables in X express the same measurement unit. Note that even though the correlation matrix supposedly handles the differentiation between the unit measurements of the variables, data standardization is highly recommended in the presence of extreme multicollinearity regardless of the selected matrix.

Finally, the basic elements of the decomposition process are the eigenvalues and the corresponding eigenvectors, which are related according to the equation:

$$Xv = \lambda v.$$

As Cheever (2020) states *"In this equation, X is an n-by-n matrix, v is a non-zero n-by-1 vector, and λ is a scalar (which may be either real or complex).Any value of λ for which this equation has a solution is known as an eigenvalue of the matrix X. It is sometimes also called the characteristic value. The vector v, which corresponds to this value is called an eigenvector"*.

As mentioned above, in PCA the PCs are oriented in directions of maximal variance of data points. In other words, the method initially aims to find the direction of a unit length vector p_1 that maximizes the variance of the score, i.e. the values that are loaded in vector t_1. This is equivalent to maximizing the function g:

$$g = t_1^\intercal t_1$$

under the constraint $p_1^\intercal p_1 = 1$ and considering that $t_1 = Xp_1$. Hence:

$$g = t_1^\intercal t_1 = p_1^\intercal X^\intercal X p_1. \tag{2.2}$$

Using Lagrange multiplier, Equation 2.2 can be transformed into:

$$g = t_1^\intercal t_1 = p_1^\intercal X^\intercal X p_1 - \lambda(p_1^\intercal p_1 - 1). \tag{2.3}$$

Taking partial derivatives:

$$\frac{\partial g}{\partial p_1} = 0 \Rightarrow \frac{\partial\{p_1^\intercal X^\intercal X p_1 - \lambda(p_1^\intercal p_1 - 1)\}}{\partial p_1} = 0 \Rightarrow$$

$$2X^\intercal X p_1 - 2\lambda p_1 = 0 \Rightarrow$$

$$(X^\intercal X - \lambda I_{p\times p})p_1 = 0 \Rightarrow$$

$$X^\intercal X p_1 = \lambda p_1. \tag{2.4}$$

From Equation 2.2 and considering Equation 2.4 :

$$t_1^\intercal t_1 = p_1^\intercal (X^\intercal X p_1) \Rightarrow$$

$$t_1^\intercal t_1 = p_1^\intercal \lambda p_1 = \lambda p_1^\intercal p_1 \Rightarrow$$

$$t_1^\intercal t_1 = \lambda. \tag{2.5}$$

Similarly, the rest requested directions of PCs can be computed, under the additional constraint of orthogonality among all of them.

Finally, it is concluded that the directions of PCs are the directions of eigenvectors of covariance matrix $X^\intercal X$. Therefore, loadings matrix P is formed by setting as columns the eigenvectors of $X^\intercal X$ and they are ordered according to the value of the respective eigenvalue. Eigenvectors with larger eigenvalues are set first. In this way, arising PCs, which consist of the columns in the XP product matrix, have a maximum variance, because their variance is equal to the respective eigenvalue, as shown in Equation 2.5.

Eventually, the mechanism of eigenvalue decomposition of a set of predictors X, where variables X_i are mean-centered, can be summarized in two steps: The creation of the covariance or correlation matrix and the computation of its eigenvectors and eigenvalues. Finally, order the eigenvalues in a declining scale and form loadings matrix P using the eigenvectors. This matrix can be used to produce scores matrix T, by setting $T = XP$.

Choosing the number of Principal Components

The major aim of PCA is dimension reduction. In other words, PCA is applied to replace the p variables-columns of an X matrix with a smaller number m of PCs, without discarding a significant amount of information (Jolliffe (2002)). Although typically p PCs can be computed, it's meaningless to work with all of them. The crucial question is how many PCs should eventually be included in the PCA model. The answer is not straightforward, as the analyst should consider a trade-off between information loss and the insertion of noise. Next, are presented the most often approaches used to determine the appropriate number of PCs:

- **Cumulative Percentage of Total Variation**
  A direct estimate of the appropriate number of PCs can be formed by the inspection of the cumulative percentage of the total variation of X, which can be explained by the inclusion of different numbers of PCs. It should be noted

that the percentage of variance explained by the i^{th} PC can be computed by the formula (Wehrens (2011)):

$$q_i = 100\frac{\lambda_i}{\sum_{j=1}^{p}\lambda_j}$$

where λ_i refers to the eigenvalue of the i_{th} PC and p the overall number of PCs. Usually, one selects the first m PCs, which absorb $80\% - 90\%$ of initial data variation (Wehrens (2011)):

$$\sum_{j=1}^{m} q_i = 100\frac{\sum_{j=1}^{m}\lambda_i}{\sum_{j=1}^{p}\lambda_j}.$$

- **Size of variances of PCs**
  This approach is also called Kaiser's Rule and is mainly applied in cases where PCs are generated by the analysis of the correlation matrix. According to Kaiser (1960), PCs are included in the PCA model as long as their variance is larger than 1. However, it should be mentioned that in Jolliffe (1972) is suggested a lower variance threshold, a value of 0.7, due to independency conditions and sampling variances. In the PCA of a covariance matrix, a sufficient threshold is considered the average value of the eigenvalues.

- **Catell's Scree Test – Scree graph**

  It is a graphic way to judge the number of PCs. On the Scree graph, also called the Scree plot, the proportion of variance explained by each PC against the rank of each PC is plotted. Usually, the curve that connects the points forms an elbow-like shape. The point located at its angle indicates the last PC to be included in the PCA model (Cattell (1966)). When the eigenvalues drop dramatically in size, an additional PC would add relatively little to the information already extracted.

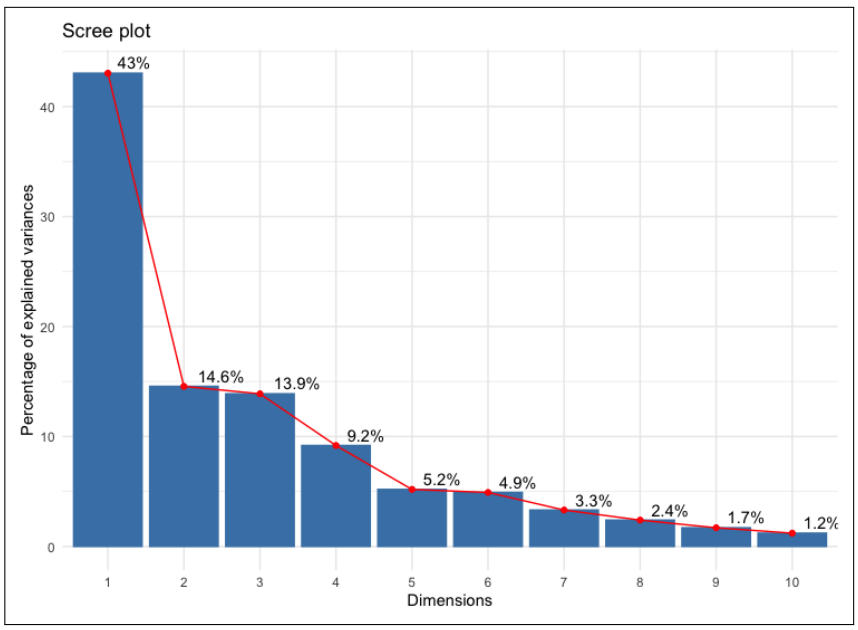


FIGURE 2.1: Catell's Scree Test

Figure 2.1 depicts a typical scree graph for a dataset with ten PCs, as well as a bar graph-like graph displaying the percentage of variation associated with each PC (value an top of each box). Based on the results, the top 5 or 6 PCs maintained the required proportion of variance from the original data ($\approx$ 85% - 90%).

There have also been proposed Cross-validation and bootstrap techniques, but they are not common due to the computational cost, especially when processing large data sets (Jolliffe (2002)).

Generally, it is recommended the use the first 2 or 3 PCs since they contain a significant amount of the original information and it is possible to visualize them in a 2- or 3-dimensional space (Figure 2.2).

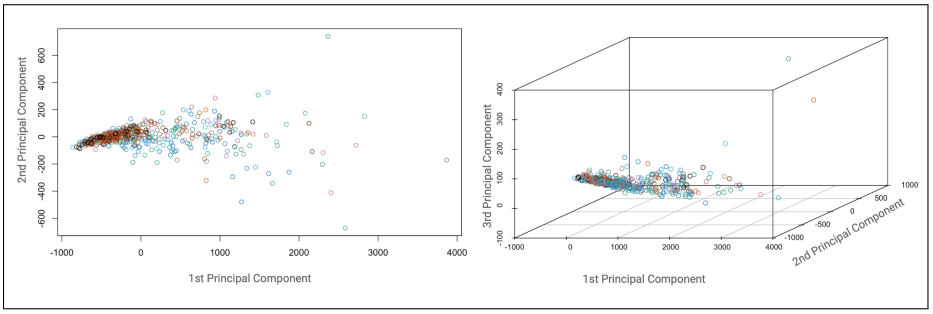


FIGURE 2.2: Visual representation of PCA on a 2D and 3D space

2.1.2 PCA constructing algorithm step by step in R

Algorithm 1 PCA implementation in R

Input: A data set consisted by a $n \times p$ matrix X, where each X_j column represents a variable.

Output: A data set consisted by a $p \times w$ matrix, where $w \leq p$ is the number of the selected PC.

Step 1: If data set is standardized then move to **Step 2**, else do:

$$\frac{x_{ij} - \bar{x}_j}{s_j}$$

where s_j indicates the standard deviation of the j^{th} variable.

```
scaled_data <- scale(data, center = TRUE, scale = TRUE)
```

Step 2: Compute covariance or correlation matrix (only correlation matrix is displayed below rounded with to two first decimals).

$$\rho(X) = \begin{bmatrix} 1 & cor(X_1, X_2) & \dots & cor(X_1, X_p) \\ cor(X_2, X_1) & 1 & \dots & cor(X_2, X_p) \\ \vdots & \vdots & & \vdots \\ cor(X_p, X_1) & cor(X_p, X_2) & \dots & 1 \end{bmatrix}$$

```
cor_matrix <- round(cor(scaled_data, method = c("pearson")),2)
```

Step 3: Compute eigenvalues and eigenvectors

$$\begin{bmatrix} A_{11} & A_{12} & \dots & A_{1n} \\ A_{21} & A_{22} & \dots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \dots & A_{nn} \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} = \lambda \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$$

```
eigen <- eigen(cor_matrix)
eigenvalues <-eigen$values
eigenvectors <- eigen$vectors
```

Step 4: Compute the Component matrix.

According to Ntotsis et al. (2020), the methodology for the construction of the components (uncorrelated vectors) is summarized below:
Let us denote by C_j the j^{th} component, λ_j the corresponding eigenvalue, and v_j the corresponding eigenvector, $j = 1, 2, ..., m$,where m represents the total number of original covariates.

Hence, C_j is defined most often with one of the following formulas; However, other variations have also been proposed (the calculation of the first is presented below)

$$C_j = \sqrt{\lambda_j}\, v_j = \sqrt{\lambda_j} \begin{bmatrix} v_{1j} \\ v_{2j} \\ \vdots \\ v_{mj} \end{bmatrix} \text{ or } C_j = -\sqrt{\lambda_j}\, v_j = -\sqrt{\lambda_j} \begin{bmatrix} v_{1j} \\ v_{2j} \\ \vdots \\ v_{mj} \end{bmatrix} \text{ or } C_j = \begin{bmatrix} v_{1j} \\ v_{2j} \\ \vdots \\ v_{mj} \end{bmatrix}$$

```
components <- eigenvectors %*% diag(sqrt(eigenvalues))
```

Step 5: Compute the Principal Component, which can be written as a linear combination between the components and the input matrix.
Each new variable Z_j, $j = 1, ..., m$ is a linear function of the components matrix and the data. We denote by $X = \{X_{ij}\}_{j=1,2,...,m}^{i=1,2,...,n}$ the matrix elements of the original data and by Z_j the new covariates, respectively.

Then, for the elements of $Z_j = (Z_{1j}, ..., Z_{nj})^\intercal$ we have:

$$Z_{ij} = \sum_{j=1}^{m} X_{ij} C_j \qquad \forall\, i = 1, ..., n, j = 1, ..., m.$$

```
pca_variables <- matrix(0, nrow(scaled_data), ncol(components))
for (i in 1:nrow(scaled_data)) {
for (j in 1:ncol(components)) {
pca_variables[i, j] <- sum(scaled_data[i, ] * components[, j])}}
```

Step 6: Select the number of optimal PC

- based on Kaiser's rule:

```
kaisers_values <- c()
for (i in 1:ncol(scaled_data)) {
if (eigenvalues[i] >= 0.99) {
kaisers_values[i] <- eigenvalues[i]}}
kaisers_comp <-
components[, c(1:NROW(kaisers_values))]
```

- based on proportion of variance explained

```
variance_decomp <-
rbind(
Total_Initial_Eigenvalues = eigenvalues,
# eigenvalues
Percentage_Variance = eigenvalues / sum(eigenvalues),
# variability explained be each component
Cumulative_Variance = cumsum(eigenvalues) / sum(eigenvalues))
variance_decomp
M <- #eigenvalue threshold that determine the number of components
var_dec_values <- c()
for (i in 1:ncol(scaled_data)) {
if (eigenvalues[i] >= M) {
var_dec_values[i] <- eigenvalues[i]}}
var_dec_comp <-
components[, 1:NROW(var_dec_values)]
```

2.1.3 Assumptions for performing PCA

When deciding to use PCA to analyse data, it must first be ensured that the data under examination must be compatible with PCA. This is necessary because PCA can only be used if the data "passes" 5 assumptions that must be met for PCA to produce a valid result. In practice, checking for these assumptions involves a few tests as well as some additional thought about the data, but it is not a tough effort.

However, it is not uncommon if one or more of these assumptions are violated when analysing data. When working with real-world data rather than textbook examples, this is not unusual. Even if the evidence contradicts certain assumptions, there is usually a way to work around it. First, consider the following five assumptions:

#1: The variables should be measured on a continuous basis (although ordinal variables are very frequently used). Continuous variables (i.e., ratio or interval variables) include revision time (measured in hours), intelligence (measured using an IQ score), exam performance (measured from 0 to 100), and so on.

#2: The available data should be suitable for DR. The existence of adequate correlations between variables in order to reduce them to a smaller number of components.

#3: All variables must be related in a linear relation. The reason for this assumption is that a PCA is based on Pearson correlation coefficients, and as such, the variables must have a linear relationship. With the use of ordinal data for variables, this assumption is somewhat relaxed (even if it shouldn't be). Although a matrix scatterplot can be used to test linearity, this is often considered overkill because a scatterplot can contain thousands of linear relationships. Thus, it is recommended a random sampling testing of this assumption between the variables.

#4: Sampling adequacy is needed, which simply means that large enough sample sizes are required for PCA to produce a reliable result. Many different guidelines have been proposed. These are primarily differentiated by whether an absolute sample size is proposed or a multiple of the number of variables in your sample is used. A prevail rule of thumb that occurs states that a minimum sample size of 100+ cases is required for sampling adequacy, which can be determined with several methods, for instance: (i) the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy for the entire data set, and (ii) the KMO measure for each individual variable.

#5: There should be no notable outliers. Outliers are significant because they can have a disproportionate impact on the outcomes.

2.1.4 Principal Component Regression

The Principal Component Regression (PCR) is a linear regression method that uses PCA and a regression step to overcome the weaknesses of Multiple Linear Regression (MLR). It is applied in case of multicollinearity between the predictors or/and in case the number of predictors is large compared to the number of available samples. Then, the X matrix can be decomposed according to PCA and, after determining the number of PCs retained in the model, low-dimension matrix T is used in MLR instead of the X matrix. As described in Wehrens (2011):

$$\begin{aligned} Y = XB + E \simeq \tilde{X}B + E' &= (TP^\intercal)B + \tilde{E} \\ &= T(P^\intercal B) + \tilde{E} = TA + \tilde{E}. \end{aligned}$$

It is obvious that PCR initially decomposes the data matrix and then replaces it with scores matrix T in a regression step. Matrix A indicates the regression coefficients to be computed. The formula for this is:

$$A = (T^\intercal T)^{-1} T^\intercal Y$$

as computed in classic Linear Regression when the Ordinary Least Squares Method is applied. It should be noted that values in the A matrix refer to scores. Matrix B with the regression coefficients that refer to the original variables of X can be computed by:

$$B = PA = P(T^\intercal T)^{-1} T^\intercal Y.$$

2.2 Partial Least Squares Method

In PCA, analysis is applied to the X data matrix. As a result, arising PCs retained in the model may not contain information related to the dependent-response variables they are next meant to predict in the regression step. In other words, valuable for prediction purposes information may be summarized in PCs that are not included in the PCA model. Partial Least Squares (PLS) is an alternative method that copes with this deficiency of PCA and, at the same time, achieves dimension reduction. As stated in Rosipal and Kramer (2005), *"It comprises of regression and classification tasks as well as dimension reduction techniques and modelling tools."* This dissertation explores PLS in both one-dimensional and multidimensional concepts (i.e., when there is only one response variable and when there are multiple response variables, respectively).

This method shares the same main idea as PCA: it forms new variables, as linear combinations of the original, which are uncorrelated. The difference is that they retain information involved both in X and Y data matrices. So, here, the aim is to generate latent variables in the direction of maximum covariance between X and Y: as stated in Wehrens (2011), *"PLS explicitly aims to construct latent variables in such a way as to capture most variance in X and Y, and to maximize the correlation between these matrices"*. However, the algorithm to achieve this goal is a bit more complicated, since both X and Y matrices are analyzed. New latent variables are generated through an iterative procedure, in every step of which is computed a set of scores vectors, a set of loadings, but also a set of weight vectors. One vector of each type refers to the X matrix and the other refers to the Y matrix. Next, follows deflation of X and Y matrices, so as to subtract the information explained by the computed components. Deflated matrices are used in the next iteration of the algorithm to generate new components. Finally, the user selects the number of components of the X matrix that will be used. Of course, their number is significantly reduced compared to the number of original variables, since valuable information is summarized in the first few PCs.

Geometrically, just like PCA, in PLS, the Y-matrix in addition to the X-matrix is projected to new spaces, the coordinate systems of which are formed by new latent variables formed by linear combinations of the original ones as mentioned before. Latent variables of the X matrix are generated so as to be orthogonal, and as a result uncorrelated, but this is not necessarily the case for latent variables of the Y matrix.

Partial Least Squares analysis of an X matrix of size $n \times p$ and a matrix Y of size $n \times k$ is:

$$X = T_\alpha P_\alpha^\intercal$$

$$Y = U_\alpha Q_\alpha^\intercal$$

where T and U are scores matrices. Again, the i^{th} columns in T and U matrices, are the coordinates of the samples in the direction of the i^{th} new latent variable, values arising from the perpendicular projection of each sample onto this direction and are measured from the origin. P and Q are loadings matrices, and the superscript $^\intercal$ indicates the transpose of a matrix. In the PLS model, loadings are vectors used in the deflation process of sequential deflated matrices generated through the algorithm. The subscript α indicates the number of latent variables that we usually compute and is equal to min(n,p,k). Since only the first few components are adequate for the following modelling purposes, the original matrices can be rewritten as (Rosipal and Kramer (2005))

$$X = T_m P_m^\intercal + E$$

$$Y = U_m Q_m^\intercal + F$$

where m is a number smaller than α and indicates the number of latent variables retained in the model. Matrices E and F contain information not explained by the first m selected Components. Eventually, the new, low-dimension matrix can be written as:

$$\tilde{X} = T_m P_m^\intercal$$

where $\tilde{X}$ indicates the approximation of X, which can be used for further modelling purposes.

2.2.1 Steps to build a PLS model

The very first step in building a PLS model is mean-centering columns in matrices X and Y so that each one has zero mean. Additionally, in cases of variables measured in different units, scaling should be considered. The reason and technique for mean-centering and scaling are the same as in the PCA model.

Concerning the computation of scores, loadings, and weight vectors, plenty of algorithms have been proposed. Some of the most known are named: The Eigenvector algorithm, by Höskuldsson (1988), Kernel algorithm for PLS introduced by Lindgren et al. (1993), Nonlinear Iterative Partial Least Squares (NIPALS) algorithm introduced by Wold (1975), Statistically Inspired Modification of the Partial Least Squares (SIMPLS) algorithm for PLS proposed by De Jong (1993), Orthogonal Projections to Latent Structures (O-PLS) proposed by Trygg and Wold (2002). The NIPALS algorithm is considered to be the most optimal of the above and thus only its documentation is displayed in the manuscript.

NIPALS Algorithm

The NIPALS algorithm constructs the PLS model's matrices in sequential steps. In every step, the X-scores vector, denoted by t and results from the projection of the X-data matrix on the direction of the new latent variable, is about to be constructed in X data space. Y-score vector, denoted by u, arises alike. These directions are defined by weight vectors w and c, respectively. Mathematically, this can be written as:

$$t = Xw$$

$$u = Yc/(c^\intercal c).$$

Directions of the w and c vectors are found so as to maximize the covariance between score vectors t and u, a value proportional to the quantity $t^\intercal u$. An additional constraint on unit length weight vectors is applied. To sum up, the problem to be solved can be written as (Rosipal and Kramer (2005)):

$$max\{cov(t,u)\} = max\{t^\intercal u\} = max\{(Xw)^\intercal Yc\} = max\{w^\intercal X^\intercal Yc\} \tag{2.6}$$

In every step, the computation of weights and scores is followed by the deflation of X and Y matrices. This process is based on p and q loadings of X and Y matrices, respectively, computed as:

$$p = X^\intercal t/(t^\intercal t)$$
$$q = Y^\intercal u/(u^\intercal u).$$

Actually, there are several variations on how to run deflation. The choice depends on the aim of PLS modelling. Wold (1975) initially proposed to deflate matrices as follows:

$$X_{new} = X_{old} - tp^\intercal$$

$$Y_{new} = Y_{old} - uq^\intercal.$$

This version of deflation of Y is used when the PLS model is built to reflect relations between matrices of variables and the algorithm is called PLS modelling. In case a PLS model is built for prediction, the algorithm is called PLS, when there is only a one-dimensional Y response variable to be predicted (univariate case), and M-PLS, when Y, like X, is multidimensional (multivariate case). These variations run deflation as follows:

$$X_{new} = X_{old} - tp^\intercal$$

$$Y_{new} = Y_{old} - btc^\intercal$$

where

$$b = u^\intercal t/(t^\intercal t). \tag{2.7}$$

When the c vector is not scaled to have a unit length (and this is most frequently the case), as shown below, b is equal to one (Höskuldsson (1988)):

$$u^\intercal t = c^\intercal Y^\intercal t/(c^\intercal c) = c^\intercal (Y^\intercal t)/(c^\intercal c) = c^\intercal c(t^\intercal t)/(c^\intercal c) = t^\intercal t \tag{2.8}$$

and from Equation 2.7 and Equation 2.8

$$b = u^\intercal t/(t^\intercal t) = t^\intercal t/(t^\intercal t) = 1.$$

In the following, the c vector is not supposed to be scaled, so b is considered equal to one. Because of this, deflation of the Y matrix becomes:

$$Y_{new} = Y_{old} - tc^\intercal.$$

Once deflation is completed new matrices X_{new} and Y_{new}, also called residual matrices, are analysed in the next step so that the next latent variable can be extracted.

2.2.2 Pseudocode for univariate PLS

With this in mind, the procedure for the PLS version can be described as follows:

Algorithm 2 Pseudocode for PLS

Input: A $n \times p$ X-matrix and a $n \times 1$ Y-matrix where each X_j column represents an explanatory variable.

Output: Vectors t, p, w and c.

Step 1: Set the vector u as the Y column, the unique vector of response variable

Step 2: Compute X weight: $w = X^\intercal u/(u^\intercal u)$

Step 3: Scale w to be unit length vector, $||w|| = 1$

Step 4: Compute X scores: $t = Xw$

Step 5: Compute Y weight: $c = Y^\intercal t/(t^\intercal t)$

Step 6: Compute X loadings: $p = X^\intercal t/(t^\intercal t)$

Step 7: Deflation process: $X_{new} = X - tp^\intercal$

Step 8: Set $X = X_{new}$ and go to step 2

The maximal number of such components that have non-zero covariance with Y is $\min(n-1, p)$, where n the number of samples and p the number of variables in X-matrix (Boulesteix and Strimmer (2006)).

The way the weight vectors are found ensures that these give the solution to the problem formulated in Equation 2.6. To prove that we can use the fact that non-deflation of the Y matrix does not influence the results. Further, let us denote as $X_i^\intercal$ the residual matrix that is going to be used for the construction of the i^{th} latent variable. For this i^{th} dimension we denote as w_{n-1} the weight vector of $(n-1)^{th}$ iteration of steps 2 to 6 before the convergence and as w_n the weight vector of n^{th} iteration before the convergence. Then, as stated in Höskuldsson (1988), the weight vector can be analysed as:

$$
\begin{aligned}
w_n &= X_i^\intercal u_{n-1}/(u_{n-1}^\intercal u_{n-1}) = \\
&= X_i^\intercal Y c_{n-1}/(u_{n-1}^\intercal u_{n-1})(c_{n-1}^\intercal c_{n-1}) = \\
&= X_i^\intercal Y Y^\intercal t_{n-1}/(u_{n-1}^\intercal u_{n-1})(c_{n-1}^\intercal c_{n-1})(t_{n-1}^\intercal t_{n-1}) = \\
&= X_i^\intercal Y Y^\intercal X_i w_{n-1}/(u_{n-1}^\intercal u_{n-1})(c_{n-1}^\intercal c_{n-1})(t_{n-1}^\intercal t_{n-1}).
\end{aligned}
$$

Considering that the i^{th} latent variable takes s iterations till convergence is achieved, then we can conclude that w_s and w_{s-1} do not differ significantly. So, by previous equation vector w_s is eigenvector of matrix $X_i^\intercal Y Y^\intercal X_i$.

Accordingly, we can find that c_s is the eigenvector of matrix $Y^\intercal X_i X_i^\intercal Y$ (Höskuldsson (1988)):

$$\begin{aligned} c_n &= Y^\intercal t_n / (t_n^\intercal t_n) = \\ &= Y^\intercal X_i w_n / (t_n^\intercal t_n) = \\ &= Y^\intercal X_i X_i^\intercal u_{n-1} / (t_n^\intercal t_n)(u_{n-1}^\intercal u_{n-1}) = \\ &= Y^\intercal X_i X_i^\intercal Y c_{n-1} / (t_n^\intercal t_n)(u_{n-1}^\intercal u_{n-1})(c_{n-1}^\intercal c_{n-1}). \end{aligned}$$

Eventually, w_s and c_s are the first eigenvectors that correspond to the largest eigenvalue of $X_i^\intercal Y Y^\intercal X_i$ and $Y^\intercal X_i X_i^\intercal Y$ matrices, respectively. Therefore, from SVD properties, these vectors maximize the quantity of interest in Equation 2.6.

Furthermore, PLS latent variables not only explain maximum covariance between X and Y but also are mutually orthogonal. The retrospective relation between residual matrices of the PLS model will help to prove it (Höskuldsson (1988)).

$$\begin{aligned} X_j &= X_{j-1} - t_{j-1} p_{j-1}^\intercal = \\ &= X_{j-1} - X_{j-1} w_{j-1} t_{j-1}^\intercal X_{j-1} / (t_{j-1}^\intercal t_{j-1}) = \\ &= X_{j-1}[I - w_{j-1} t_{j-1}^\intercal X_{j-1} / (t_{j-1}^\intercal t_{j-1})] = \\ &= [X_{j-2} - t_{j-2} p_{j-2}^\intercal][I - w_{j-1} t_{j-1}^\intercal X_{j-1} / (t_{j-1}^\intercal t_{j-1})] = \\ &= [X_{j-2} - X_{j-2} w_{j-2} t_{j-2}^\intercal X_{j-2} / (t_{j-2}^\intercal t_{j-2})][I - w_{j-1} t_{j-1}^\intercal X_{j-1} / (t_{j-1}^\intercal t_{j-1})] = ... \end{aligned}$$

$$= [X_i - X_i w_i t_i^\intercal X_i / (t_i^\intercal t_i)]...[I - w_{j-2} t_{j-2}^\intercal / (t_{j-2}^\intercal t_{j-2})][I - w_{j-1} t_{j-1}^\intercal X_{j-1} / (t_{j-1}^\intercal t_{j-1})].$$

Next, the proof of orthogonality between scores vectors follows:

Let indices i and j denote now two different directions of extracted latent variables (suppose $i < j$). From the retroactive relation above, we can write (Höskuldsson (1988)):

$$\begin{aligned} X_j = [X_i - X_i w_i t_i^\intercal X_i / (t_i^\intercal t_i)]_{...}[I - w_{j-2} t_{j-2}^\intercal / (t_{j-2}^\intercal t_{j-2})] \\ [I - w_{j-1} t_{j-1}^\intercal X_{j-1} / (t_{j-1}^\intercal t_{j-1})] = \\ = [X_i - X_i w_i t_i^\intercal X_i / (t_i^\intercal t_i)] Z \end{aligned}$$

where Z some matrix.

Further,

$$\begin{aligned} t_i^\intercal X_j &= t_i^\intercal [X_i - X_i w_i t_i^\intercal X_i/(t_i^\intercal t_i)]Z = \\ &= t_i^\intercal X_i - (t_i^\intercal X_i w_i t_i^\intercal X_i)/(t_i^\intercal t_i) = \\ &= t_i^\intercal X_i - (t_i^\intercal t_i) t_i^\intercal X_i/(t_i^\intercal t_i) = \\ &= t_i^\intercal X_i - t_i^\intercal X_i = 0. \end{aligned} \tag{2.9}$$

Consequently,

$$t_i^\intercal t_j = t_i^\intercal X_j w_j = 0.$$

This means that score vectors are mutually orthogonal and as a result uncorrelated, a very significant property.

When the whole process of extracting latent variables is completed, the involved t scores vectors, p loadings vectors, w and c weight vectors are combined as column-vectors and form respectively T scores matrix, P loadings matrix and weight matrices W and C.

However, frequently, for interpretation purposes, another matrix is being computed:

$$R = W(P^\intercal W)^{-1}.$$

The need for R arises because derived weight scores do not refer to the original matrix X and its original variables, but to the sequential deflated matrices X_i. On the contrary, each column vector in the R matrix expresses the weights of the original variables of X at the corresponding dimension (Wehrens (2011)). Algebraically, is the generalized inverse of matrix $P^\intercal$, which is singular, and it is

$$T = XR.$$

2.2.3 Pseudocode for multivariate PLS

M-PLS is a more computationally complex procedure, compared to PLS, and can be described by the following pseudocode:

Algorithm 3 Pseudocode for M-PLS

Input: A $n \times p$ X-matrix and a $n \times k$ Y-matrix where each X_j column represents an explanatory variable and Y_j represents a response variable.

Output: Vectors t, p, w and c.

Step 1: Set the vector u as the first or any other column of Y

Step 2: Compute X weight: $w = X^\intercal u/(u^\intercal u)$

Step 3: Scale w to be unit length vector, $||w|| = 1$

Step 4: Compute X scores: $t = Xw$

Step 5: Compute Y weight: $c = Y^\intercal t/(t^\intercal t)$

Step 6: Update u scores vector: $u = Yc/(c^\intercal c)$

Step 7: Test convergence of ratio $v = ||t_{old} - t_{new}||/||t_{new}||$

- If $v > \epsilon$, go to step 2 (where ϵ set to a number between $(10^{-8}, 10^{-6})$ for instance)
- If $v < \epsilon$, go to step 8

Step 8: Compute X loadings: $p = X^\intercal t/(t^\intercal t)$

Step 9: Deflation process: $X_{new} = X - tp^\intercal$ and $Y_{new} = Y - tc^\intercal$

Step 10: Set $X = X_{new}$ and $Y = Y_{new}$ and go to step 2

The way t scores are derived implies that they also contain information about Y. As a result, they are also good predictors of Y (and that is the reason that deflation of Y matrix is done by subtracting $tc^\intercal = tt^\intercal Y/(t^\intercal t)$, where $t^\intercal Y/(t^\intercal t)$ is the OLS estimate v of coefficient in regression $Y = tv$).

At this point, it should be mentioned that in Wold et al. (2001) it is supported that deflation of Y matrix is optional, since, as it is stated there *"the results are equivalent with or without Y-deflation"*.

2.2.4 Partial Least Squares Regression

Partial Least Squares Regression (PLSR) is a linear regression method that uses PLS, as a method of extracting latent variables and a regression step. It is used when Multiple Linear Regression is impossible to give a solution to a regression problem or its solution is not stable. Typical situations include collinear predictors and/or the need for dimension reduction. Then, PLS or M-PLS can be applied to X and Y matrices, depending on the dimension of Y. The PLS model with the selected components is then used in regression:

$$\begin{aligned} Y = XB + E \simeq \tilde{X}B + \tilde{E} &= (TP^\intercal)B + \tilde{E} \\ &= T(P^\intercal B) + \tilde{E} = TA + \tilde{E}. \end{aligned}$$

The above regression scheme, as presented in Wehrens (2011), is identical to PCR. The difference lies in the computation of scores, which takes into account the response variable(s). Regression coefficients are again computed using the Ordinary Least Squares Method.

$$A = (T^\intercal T)^{-1} T^\intercal Y.$$

In PLS A and $\tilde{E}$ are column vectors, while in M-PLS they are matrices, where the number of their columns is equal to the number of response variables, exactly like MLR. Matrix B with the regression coefficients that refer to the original variables of X can be computed by the fact that the inverse matrix of $P^\intercal$ is R:

$$B = RA = R(T^\intercal T)^{-1}T^\intercal Y.$$

Note: In NIPALS, algorithm loading vectors p are not mutually orthogonal as scores are. However, it is interesting that in Martens and Naes (1989) it has been proposed another algorithm that generates orthogonal loadings instead, and it is shown that regression coefficients are the same as those resulting from PLSR (Helland (1988)).

2.3 Limits and extensions of PCA and PLS in the modelling process

Researchers must be conscious of the DRT they choose, which must be selected based on data characteristics and peculiarities in addition to the desired outcome (i.e., the actual purpose) of the analysis. PCA is an unsupervised technique meaning that only X-matrix is analysed, while PLS is considered supervised since the response variable affects the formulation process of the components.

When in PCA, columns in matrix T are independent. As a result, the matrix $T^\intercal T$ is numerically stable, and therefore, the consequences of collinearity in estimating regression coefficients are eliminated. Furthermore, since dimension reduction is achieved by selecting only m out of p PCs to retain in the PCA model, the complexity of the final regression model is significantly optimized. PCs can be used to create informative visualizations of multivariate datasets and in addition, they can be combined with other (supervised) methods, where PCA can be used to pre-process the dataset. Unlike many unsupervised methods, PCA can be applied even when the number of available samples is lower than the number of regressors in a dataset. However, like the majority of feature extraction techniques, PCA arises with some critical issues. Most importantly, the coefficients arising from the regression step of PCA refer to new variables. Therefore, the interpretability of the original variables is lost in the process of creating the new PCs. Additionally, the transformation of original variables to PCs can cause the loss of vital information, especially for prediction purposes. Lastly, the selection of the optimal PCs to be retained for the modelling process model is subjective, meaning the probability of model overfitting or underfitting is unpredicted.

PLS and PCA when modelling, share the same main philosophy as regression methods; therefore they present similarities when compared. As shown, utilized score vectors in PLS are mutually orthogonal and thus uncorrelated. As a result, they can replace correlated predictors in regression and they can effectively lead to the estimation of coefficients. Because of the mechanism that generates the T matrix, the information compressed in it is directly related to response variables. This means that PLS can also deal with noisy data. Additionally, when concerning big data analysis without collinear variables, dimension reduction achieved by PLS can beneficially reduce model complexity. In that sense, PLS (*PCA, and other DRTs*) are helpful when graphical representations are used to get a "big picture" of the data by giving a better understanding of the structure. PLS successfully deals with the "small n large p" problem, a situation that other supervised methods cannot overcome. This situation is very common in regression analysis of biological, chemical, and other scientific problems in which PLS consists of one of the optimal choices. Furthermore, a large number of predictors are also associated with a phenomenon called over-fitting. PLS has been proven to be able to handle multivariate modelling while performing variable extraction. Another beneficial characteristic of PLS is that

in the case of a full rank X-matrix, the PLS model that includes as many latent variables as columns in X gives an identical solution as the MLR model. However, in the case of correlated original predictors, as is most commonly the case, MLR regression predictors are misleading due to multicollinearity. On the contrary, PLSR regression coefficients are shrunk estimates and thus more robust, leading to better predictions. A multidimensional Y matrix is analyzed in different ways in two methods. In MLR, a linear regression model is produced for each response variable and the estimates of the regression coefficients are different in the generated models. In PLS, in the case of correlated responses, the variation M-PLS can be applied and one regression model is produced. Hence, regression coefficients are common for all Y variables. Besides the fact that in this way the analysis is completed very fast, the relations between the response variables play a significant role in the definition of regression coefficients. However, in cases there is no correlation between response variables, individual PLS models are a more appropriate choice.

The main drawback of the PLS method is that the regression coefficients estimated by the regression step of PLS need an extra process, so as to refer to the original variables. Additionally, the magnitude of summarized information, eventually used for predictions may depend on the user and the interpretation of methods that help choose the retaining components.

PCA and PLS are dimension reduction techniques based on the same main idea: the aim is to construct latent variables that summarize as much as possible data information and achieve dimension reduction by using the most informative of them. Despite their similarities, methods differ in the following:

- PCA achieves its purpose with a simple one-step algorithm and produces elements, meaning scores, and loadings, which refer to the original variables. On the other hand, PLS make use of iterative procedures, so scores, weights, and loadings refer to sequential deflated matrices, which impede their direct interpretation.

- Furthermore, *"in PLS dimension reduction and regression is performed simultaneously* as referred in Yeniay and Goktas (2002). In contrast, the implementation of the low-dimension matrix, resulting from the PCA of an X-matrix, in a regression scheme is a different step.

- Technically, they differ in the optimization problem they aim to solve in order to extract these latent variables. PCA derives variables by maximizing the information of the X-matrix that is explained, while PLS maximizes the covariance of X and Y matrices that is explained.

- Their main difference when occupied in regression problems is that PLS involves also information in Y to model the data (supervised method), while PCA is independent of responses (unsupervised method). There is no guarantee that the directions that best explain the predictors will also be the best directions to use for predicting the response. However, Artemiou and Li (2009) proved that when PC are used as predictors in a regression, then:

$$P\big(cor^2(Y, Z_k) > cor^2(Y, Z_n)\big) > 1/2 \qquad \forall k < n.$$

- The next statement is directly related to this difference: *"Because PLS components are developed as latent variables possessing a high correlation with Y, the optimum number of PLS components is usually smaller than the optimum number of*

PCA components in PCR" (Varmuza and Filzmoser (2009)). It is expected that the PLSR model will perform better, because it includes information coming from the overall system of variables that are being modelled, and not from a part of it, meaning the part of the system that defines matrix X. This has the advantage of fewer factors to interpret and minimization of computational cost. In Yeniay and Goktas (2002) it has been proven that PLSR models are most parsimonious and have higher predictive accuracy.

- When multivariate Y is about to be predicted, PCR will run multiple regression steps, one for the prediction of each response variable, using the same Principal Components for all. On the contrary, PLS has a variation, M-PLS, that is appropriate for dependent response variables. In this case, the same latent variables will be used for simultaneous prediction of them. The way they are generated implies that relations between Y responses are taken into account, leading to a better illustration of the whole investigated system. However, in the case of independent variables, separate PLSR models perform better; "*A single PLSR model tends to have many components and be difficult to interpret*" (Wold et al. (2001)).

- Generally, PCR and PLSR result in different regression coefficients. However, when adding components, models tend to become more similar (Boulesteix and Strimmer (2006)).

2.4 Review of regularization methods

In statistics, econometrics, and machine learning, among other fields, regularization methods are considered optimal for parsimonious model creation when an immense number of variables are involved. The use of such methods addresses the problem of model over-fitting by imposing a low predictor coefficient value when it is sparse - and by expansion can be exploited as variable selection criteria - and secondly can sustain the significant estimates in the presence of multicollinearity.

Regularization is a collection of techniques that can help avoid overfitting in the training set of statistical modelling and convolutional neural networks, improving the accuracy of deep learning models when they are fed entirely new data from the problem domain. Some of the most common regularization techniques are the Manhattan and Euclidean distances -L^1 and L^2, respectively; dropout, early stopping, and data augmentation.

A good model has the ability to generalize well from the training data to any data from the problem domain; this allows it to make good predictions on data that the model has never seen before. To define generalization, consider how well the model has learned to apply concepts to any data rather than just the data it was trained on during the training process. On the other hand, if the model is not generalized, an overfitting problem arises. Overfitting occurs when the model performs well on training data but fails when applied to testing data. It even detects noise and fluctuations in the training data and learns from them.

The most prevalent types of regularization are Ridge, Lasso, and their aggregation, Elastic Net Regularization (ENR). These techniques are based on norms and are particularly useful tools to mitigate the issue of multicollinearity since they are subject to the premise that smaller weights result in simpler models, which helps to avoid overfitting. To obtain a smaller matrix, these techniques include a "regularization term" in addition to the loss in order to obtain the cost function. For the use of

regularizations, two tuning parameters are computed. Firstly, the mixing parameter $\alpha \in [0, 1]$, which combats over-fitting by constraining the size of the weights. Secondly, the non-negative regularization parameter λ minimizes the prediction error (*MSE*) by controlling the model's regularization magnitude.

2.4.1 Tikhonov regularization

Ridge, which was developed by Tikhonov (1943, 1963), manages to shrink the model's complexity while preserving all variables involved by minimizing the coefficients of the insignificant variables (see also Perez-Melo and Golam-Kimbia (2020)). When in Ridge, $\alpha = \alpha_r = 0, \lambda = \lambda_r$ and the penalty function for the β_j coefficient of the j_{th} variable can be expressed:

$$p_{\alpha_r,\lambda_r}(|\beta_j|) = \lambda_r \times \beta_j^2. \tag{2.10}$$

2.4.2 Lasso regularization

On the contrary, Lasso, initially introduced in geophysics but popularized in statistics by Tibshirani (1996), manages to shrink the model's complexity by setting equal to zero all the insignificant coefficients and by dropping the corresponding variables. Therefore, it can also act as a variable selection technique that makes the model more interpretable. When in Lasso, $\alpha = \alpha_l = 1, \lambda = \lambda_l$, and the penalty function for the β_j coefficient can be expressed:

$$p_{\alpha_l,\lambda_l}(|\beta_j|) = \lambda_l \times |\beta_j|. \tag{2.11}$$

2.4.3 Elastic Net regularization

Ridge regression tends to shrink the high collinear coefficients towards each other, while Lasso picks one over the other. To manage both simultaneously, the ENR was developed as a compromise between the two, in an attempt to shrink and do a sparse selection simultaneously by mixing Lasso's and Ridge's penalties (Hastie et al. (2001)). The EN linearly combines two L^p metrics and, more precisely, the Manhattan and Euclidean distances - L^1 and L^2 penalties respectively, of the Lasso and Ridge methods (Zou and Hastie (2005)). This capability allows tuning both α and λ parameters at the same time. Tuning parameter $\alpha = \alpha_{en} \in [0, 1]$ and when in ranges endpoints, then Ridge and Lasso's regularizations arise respectively. In the case of Elastic Net tuning parameter λ is denoted as λ_{en}, while the corresponding penalty function for the β_j coefficient can be expressed as:

$$p_{\alpha_{en},\lambda_{en}}(|\beta_j|) = \lambda_{en} \times \left(\frac{1 - \alpha_{en}}{2} \beta_j^2 + \alpha_{en} |\beta_j| \right). \tag{2.12}$$

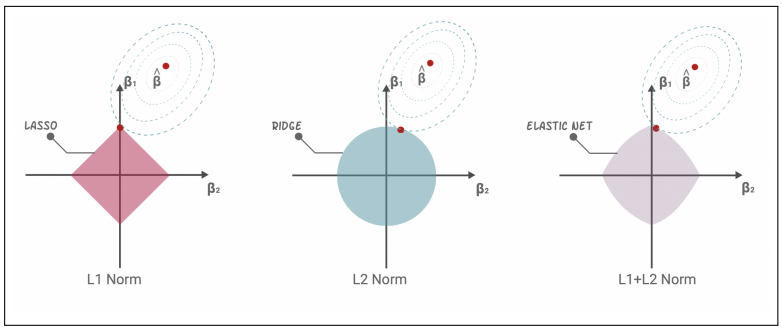


FIGURE 2.3: Geometric projection of L1 and L2 Norms' penalty terms in the space of the model parameters

In Figure 2.3 $\hat{\beta}$ is the OLS solution in all three graphs. A solid shape represents the constraint region: the square represents Lasso; the circle represents Ridge, and the rectangle with curved edges represents the Elastic Net regression.

The ellipses represent the error contours. The constraint regions will include $\hat{\beta}$ if the penalty is small or zero (i.e., $\lambda = 0$). The ellipses centred on $\hat{\beta}$ represent constant RSS regions, and values on a single ellipse have the same value as RSS. The RSS increases as we move away from the OLS coefficient estimates.

The Lasso, Ridge, and Elastic Net regression coefficient estimates are represented by red points where the ellipse touches the constraint region. When the constraint region is a square, the ellipse will intersect it on one of its axes. A constraint region that is a circle, on the other hand, will almost never touch the ellipse at an axis (and none of the coefficients will be zero). Elastic Net is a Lasso-Ridge hybrid regularization, and its ellipses will not touch the constraint region at the axis but may come very close. However, the accuracy of one model based on test data is not guaranteed to be greater than the accuracy of the other model based on test data. Hence, their utilization is not so frequent. The shrinkage of the three models varies significantly: The coefficients in ridge regression are reduced by the same proportion, whereas in lasso regression, the coefficients are shrunk towards zero by a constant amount ($\lambda/2$). Any coefficient less than $\lambda/2$ is set to zero. An Elastic Net will fall somewhere in the middle.

The disadvantage of this method is that it can be computationally time-consuming due to all the possible α_{en} values (Liu and Li (2017)) that need to be considered, especially when the case requires the procedure to be repeated as many times as the number of variables involved. In order to resolve this issue, along with the ones arising from standard measures of multicollinearity, a new robust criterion will be proposed in Section 6 as a specialized advanced regularization method.

2.4.4 Data augmentation

The objective of DRT is not only to reduce the dimension of the problem under study, but also to correct the dimension, i.e., to seek the dimension that leads to the best solution to the problem. Data augmentation (DA) can be considered as a dimension "correction" technique that reduces overfitting by "collecting" (producing) new data. Data augmentation is a regularization technique that is commonly used for image classification, signal processing, time series, and speech recognition, among others. The DA is closely associated with oversampling when it comes to data analysis. Rearranging components of real data to create synthetic signals is a frequent DA technique in signal processing, while in time series analysis, block bootstrap augmentation is a preferred DA approach.

In image recognition, DA is implemented for artificial generation of additional data from the existing training data by making minor changes to the image, such as rotation, flipping, cropping, or blurring a few pixels, and this process generates more and more data. The model variance is reduced by using this regularization technique, which reduces the regularization error. DA can usually be found in the area of machine learning the training purposes. Augmentation has the potential to increase the amount of relevant data in a dataset and hence it is an optimal way to "feed" information to the neural network in order to train it. The most common augmentations have been presented below with an image illustration. In Figure 2.4 the implementation of these augmentations takes place on two images of a cat.

- **i. Flip**
  Images can be flipped horizontally and vertically. Vertical flips are not supported by all frameworks. A vertical flip, on the other hand, is equivalent to rotating an image by 180 degrees and then performing a horizontal flip.

- **ii. Scale**
  The image can be scaled inward or outward. The final image size will be larger than the original image size as you scale outward. Most image frameworks extract a section of the new image that is the same size as the original image.

- **iii. Crop**
  Unlike scale, in this case, sampling of a section of the original image takes place at random. This section is then resized to the original image size. This is commonly referred to as random cropping. The differentiation between approach and scaling can be seen in Figure 2.4.

- **iv. Rotation**
  One important aspect of this operation to keep in mind is that image dimensions may not be preserved after rotation. If the image is square, rotating it at right angles will keep the image size intact. If it's a rectangle, rotating it by 180 degrees will keep the size the same. The final image size will change as the image is rotated at finer angles.

- **v. Translation**
  The translation is simply moving the image along the horizontal and/or the vertical axis. This approach is a bit more complex than the others, since if the input image is the original (unedited) one, then it is impossible to move between the axes. The most simple way to resolve this issue is to remove the background of the image and only keep the target point (in this scenario, the cat), and train the edited image. In Figure 2.4 the assumption that the input image has a white background beyond its boundary and translates it accordingly. Because most objects can be found almost anywhere in the image, this method of augmentation is extremely useful. This forces the neural network to search in every direction.

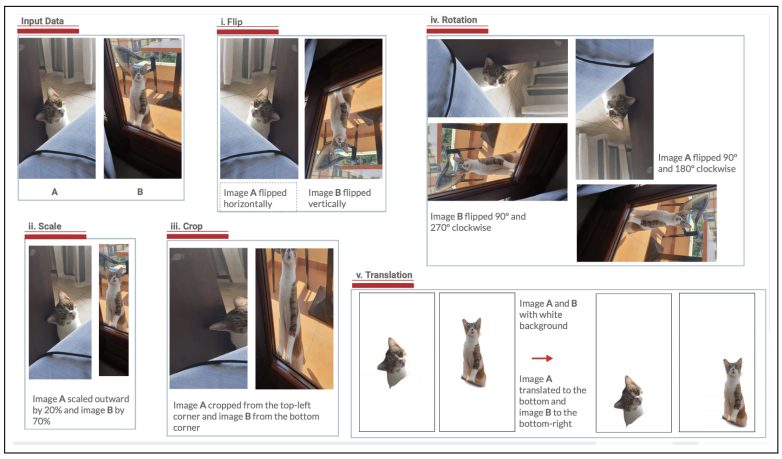


FIGURE 2.4: Image data augmentations

2.4.5 Early stopping

It is a cross-validation strategy in which one part of the training set is used as a validation set and the model's performance is measured against this set. As a result, if the model's performance on this validation set deteriorates, training on the model is halted immediately.

The basic idea behind this technique is that after each iteration of fitting a neural network on training data, the model is evaluated on unseen data or the validation set. So, if performance on this validation set decreases or remains constant over time, the model training process is terminated. This technique is used to address the issue of overfitting in the model.

2.4.6 Dropout

Another popular regularization technique that is frequently used in neural networks is dropout. It essentially means that during neural network training, randomly selected neurons are turned off or "dropped." In DR, variational dropout has been proposed as a feature selection technique because it works by assigning ranks to features. This procedure generates a "network" that contains only the important features, i.e., those with a low dropout rate.

Because in dropout, the dropping happens randomly for different units on each layer, each iteration can be viewed as a different model. This means that the error would be the average of all model errors. As a result, averaging errors from different models, particularly if those errors are uncorrelated, reduces overall errors. In the worst-case scenario, where errors are perfectly correlated, averaging across all models will be useless; however, it is known that errors have some degree of no correlation in practice. As a result, generalization errors are always reduced.

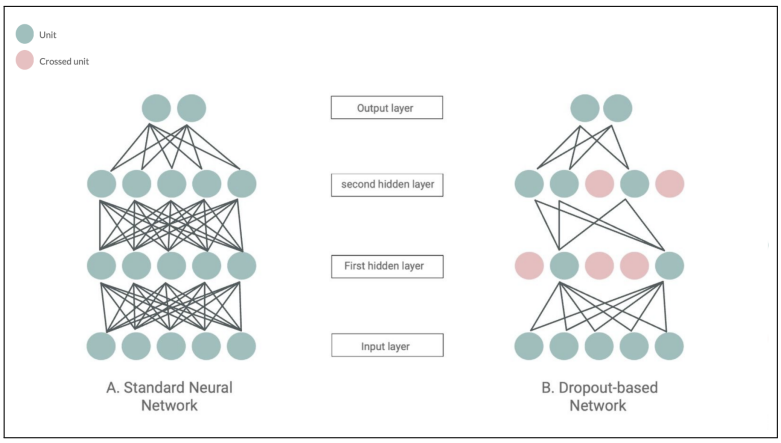


FIGURE 2.5: Dropout regularization for dimensionality reduction

Figure 2.5 displays how a dropout mechanism can be used as a feature selection DRT in a neural network-like graph. (A) illustrates a standard neural network with two hidden layers, and (B) illustrates an example of a thinned network produced by dropout, with the red units to signify the dropped variables.

Chapter 3

Model Assessment Criteria

3.1 A guide to evaluate the DRT models

In order to assess the models that the aforementioned methods (among others) generate, several criteria and indices have been developed. In this section the discussion about the model assessment criteria that were utilized throughout this thesis, takes place.

The selection of the optimal number of latent variables to retain in a PLS model is determined on the basis of the following criteria:

- Wold's R criterion: It is a criterion specially designed to evaluate PLSR models by comparing the contribution of a new extracted variable with the previous one, to the predictive ability of the model. For this purpose, a cross-validation technique is involved to compute the Predicted Error Sum of Squares (PRESS) statistic and WR ratio as follows (Li et al. (2002)):

$$WR = \frac{PRESS(m+1)}{PRESS(m)}$$

  where m denotes the number of retained latent variables in the model. The inclusion of the latent variable that makes WR greater than one, terminates the construction algorithm and the production of new latent variables. The first m of them are then included in the model.

- Adjusted Wold's R criterion: In this permutation of Wold's R criterion the ratio WR is compared to the values 0.90 ($WR_{adj}^{0.90}$) and 0.95 ($WR_{adj}^{0.95}$) rather than 1, as in the original version. As it has been proven in Li et al. (2002), these variations give better results due to sample variability.

In many cases, when researchers deal with high dimensional datasets, variable selection leads up to the construction of a PCA/PLS model, in order to remove insignificant variables at a preparatory level. As a result, the production of sets of models that differ in the number of predictors they arise from and also differ in terms of complexity occurs. The selection of the optimal model can emerge from various model selection criteria (Faraway (2002)). The most frequently utilized criteria that one can use when in PCA, PLS, and similar techniques have been documented below:

3.1.1 Coefficient of Determination – R^2

It expresses the percentage of the explained variability in the response variable and it is computed by:

$$R_p^2 = \frac{\sum_i^n (\hat{y}_i - \bar{y})^2}{\sum_i^n (y_i - \bar{y})^2}$$

where n refers to the number of available observations and p to the number of retained components. It varies from zero to one, and higher values indicate more sufficient model performance.

3.1.2 Adjusted Coefficient of Determination – $\mathbf{R}^2_{adj}$

It is a modification of the R^2 criterion that penalizes models of higher complexity. It is computed by:

$$R^2_{adj,p} = 1 - \frac{(n-1)}{(n-p-1)}(1 - R^2)$$

3.1.3 Akaike Information Criterion – AIC

This criterion can be considered as the relative amount of information lost by the candidate model: the less information lost, the higher the model's quality. In other words, AIC approximates the quality of a candidate model relative to each of the other candidate models for the data. As mentioned above, the task is accomplished by combining a criterion that minimizes the loss of information with a maximum likelihood estimation method (Akaike (1974)). More specifically, AIC is based on the log-likelihood function and is defined as:

$$\text{AIC}_p = -2(\textit{maximum log-likelihood}) + 2p$$

where p represents the dimension of the vector-parameter θ. The optimal model is the one with the lowest AIC value.

3.1.4 Bayesian Information Criterion – BIC

BIC is a model identification procedure based on information theory but set within a Bayesian context. It is an evaluation criterion for models estimated by using the maximum likelihood method. BIC can be considered as an estimate of a function of the posterior probability of a model being true, under a certain Bayesian setup, so that a lower BIC means that a model is considered to be more likely to be the true model (Schwarz (1978)). BIC is given by

$$\text{BIC}_p = -2(\textit{maximum log-likelihood}) + p \log n$$

where p represents the dimension of the vector-parameter θ and n is the number of observations.

3.1.5 Modified Divergence Information Criterion – MDIC

The Divergence Information Criterion (DIC) proposed by Mattheou et al. (2009) constitutes a modelling generalization of AIC, based on the Basu, Harris, Hjort, and Jones (BHHJ) divergence measure (Basu et al. (1998)). The DIC family of procedures,

like AIC, is an asymptotic approximation as the sample size increases and offers an alternative based on the so-called divergence measures (Toma (2014)). Let us consider the Modified Divergence Information Criterion (MDIC), a modification of the DIC proposed by Mantalos et al. (2010). MDIC can be viewed as an approximation of the expected overall discrepancy, which based on the BHHJ measure, evaluates the distance between the true and the fitted models. If the model with the smallest estimator of the expected overall discrepancy is chosen, then it is possible to end up with a model with an unnecessarily large number of variables. Thus, the Modified Divergence Information Criterion is a criterion comparable to the AIC. The MDIC formula is given as:

$$\text{MDIC}_p = nMQ_{\hat{\theta}} + (2\pi)^{-\frac{\alpha}{2}}(1+\alpha)^{2+\frac{p}{2}}p$$

where, for $f_\theta(\cdot)$ being the (candidate) model

- p is the order of the model or the number of variables involved,
- $$MQ_{\hat{\theta}} = -\big[(1+\frac{1}{\alpha})\frac{1}{n}\sum_{i=1}^{n}(f_{\hat{\theta}}(x_i))^{\alpha}\big],$$
- $\hat{\theta}$ is a consistent and asymptotically normal estimator of the parameter vector θ, and
- $\alpha \in (0,1)$ is the positive index, often chosen to be equal to 0.25.

3.1.6 Root Mean Square Error of Cross Validation – RMSECV

This measure involves cross validation to give an estimation of the variation/ divergence of the predicted values from the true values of unseen observations, in lack of available data that could be used as a test set. The criterion uses the cross-validation approach and its value is computed as:

$$RMSECV = \sqrt{\frac{\sum_j \frac{\sum_i (y_{ij}-\hat{y}_{ij})^2}{N_j}}{k}}$$

where $\hat{y}_{ij}$ is the estimation of y_{ij}, N_j is the number of observations in the j^{th} fold and k is the number of folds in cross-validation procedure. Lower RMSECV values indicate better predictive capacity of the compared models.

3.1.7 Stepwise Regression – step

Stepwise regression is a method of fitting regression models in which the selection of predictive variables is done automatically. Each step considers a variable for addition to or subtraction from the set of explanatory variables based on some predetermined criterion. This is typically done in the form of (i) forward selection, which entails starting with no variables in the model, testing the addition of each variable using a chosen model fit criterion, adding the variable (if any) whose inclusion gives the most statistically significant improvement in the fit, and repeating this process until none improves the model statistically significantly, (ii) backward elimination, which entails starting with all candidate variables, testing the deletion of each variable against a chosen model fit criterion, deleting the variable (if any) whose loss results in the least statistically significant deterioration of the model fit, and repeating

this process until no more variables can be deleted without a statistically significant loss of fit, and so on and (iii) bidirectional elimination, a combination of the above, testing for variables to be included or excluded at each step.

The common practise of fitting the final selected model followed by reporting estimates and confidence intervals without adjusting them to account for the model building process has prompted calls to abandon stepwise model building entirely or, at the very least, to ensure model uncertainty is correctly reflected. Alternative model selection techniques, such as, R^2_{adj} (stepR^2_{adj}), AIC (stepAIC), BIC (stepBIC), DIC (stepDIC), etc., can be utilized from stepwise regression.

3.1.8 Correlation-Based Feature Selection – CFS

A correlation-based Feature (variable) Selection, denoted by CFS (Hall (1999)), can be used as an alternative approach for DR. The CFS is a measure that evaluates subsets of variables on the basis of the following Hall's hypothesis:

An optimal variable subset includes uncorrelated independent variables and simultaneously high correlations between each variable with the dependent variable. If such correlations are available, then the merit of a variable subset S consisting of N variables is defined as:

$$\text{Merit}_{S_N} = \frac{N\overline{r_{YX_i}}}{\sqrt{N + N(N-1)\overline{r_{X_iX_i}}}} \tag{3.1}$$

where $Merit_{S_N}$ is the correlation between the summed independent variables and the dependent variable, N is the number of variables, $\overline{r_{YX_i}}$ is the average of the correlations between the independent variables and the dependent variable, and $\overline{r_{X_iX_i}}$ is the average inter-correlation between the independent variables. Hall presented a backward elimination procedure, with the use of Equation 3.1 in order to choose a subset. The full set of variables is evaluated with Equation 3.1, which, in fact, is the Pearson's correlation coefficient with standardized variables. Then, a variable is temporarily removed and the set of variables is evaluated with the aforementioned equation. If the subset scores are higher than the set before, then the variable is permanently removed. Otherwise, it is reinstated. The process continues until each variable is removed once and the effect of its removal is measured. The process stops when no subset scores are higher than those of the original set.

Chapter 4

On the Modelling of Pension Expenditures in Europe[1]

4.1 Definition of the data framework

The purpose of this work is to identify the appropriate variables and model the Public Pension Expenditures as percentage of GDP (PPE), -which will be addressed as Pension Expenditures or simply as Expenditures in the rest of this manuscript, of various European countries. As the Organisation for Economic Co-operation and Development (OECD) (2020) states, *"Pension Expenditures, also named pension spending, is defined by all cash expenditures (including lump-sum payments) on old-age and survivors pensions. Old-age cash benefits provide an income for persons retired from the labour market or guarantee incomes when a person has reached a standard pensionable age or fulfilled the necessary contributory requirements. This category also includes early retirement pensions: pensions paid before the beneficiary has reached the standard pensionable age relevant to the program. It excludes programmes concerning early retirement for labour market reasons. Old-age pensions includes supplements for dependants paid to old-age pensioners with dependants under old-age cash benefits. Old age also include social expenditures on services for the elderly people, services such as day care and rehabilitation services, home-help services and other benefits in kind. It also includes expenditures on the provision of residential care in an institution. This indicator is measured in percentages of GDP broken down by public and private sector"*.

There are plenty of works in the literature concerning the Pension Expenditures analysis, most of which are focusing on a single country or a few variables of importance. de La Fuente (2015) analysed the pension system of Spain as a function of workers Social Security contribution histories, while Karam et al. (2010) studied and analysed the macroeconomic effects of public pension reforms. Marcinkiewicz and Chybalski (2014) discussed Pension Expenditures as one of the main indicators of pension system sustainability; proposed a model based on GDP and old-age dependency ratio, and applied the resulted model to countries with very different population structures. The same authors, later Marcinkiewicz and Chybalski (2016), suggested a new typology of pension regimes between OECD countries. The interested reader may look at Lachowska and Myck (2018), Bonoli and Shinkawa (2005), Franco et al. (2006) and Bonoli (2003) for additional information and results concerning Expenditures.

[1]The results of this Chapter have been published in:
i. Ntotsis, K., Papamichail, M., Hatzopoulos, P. and Karagrigoriou, A.: On the Modelling of Pension Expenditures in Europe, Communications in Statistics: Case Studies, Data Analysis and Applications, 6(1), 50–68, 2020.
ii. Ntotsis, K., Papamichail, M., Hatzopoulos, P. and Karagrigoriou, A.: On the Multivariate Modeling of Pension Benefits, The European Actuary, Issue No.23, 14–19, 2020.

In this work we rely on techniques including PCA and Generalized Linear Models (GLM) (McCullagh and Nelder (1989)) for the modelling PPE by identifying the appropriate set of variables from a long list of possible explanatory variables which likely act on and affect the Expenditures. For relevant approaches one can refer to (Barr (2006), Farrell (2001), Hickman (1968) and Homburg (2000)).

4.2 Preference Data

The modelling of Pension Expenditures, according to the relevant theory (Barr (2006), Diamonds (2001), Farrell (2017), Hickman (1968), Holzmann (2009), Homburg (2000), Samuelson (1958), Schneider (2005)) could be based on a number of explanatory variables. For this work 20 European countries were selected and a total of 20 explanatory variables which are most likely related and possibly affect either directly or indirectly expenditures have been chosen based on the completeness of available data collected from Knoema (2022), OECD (2022) and Statistical Office of the European Communities (Eurostat) (2019). The data which are annual, cover the period 2001 to 2015. Note that at the time of this work the data for 2016 and 2017 were not fully available. Based on the available data, in addition to the Overall dataset (2001-2015), three individual datasets were created corresponding to the time-periods 2001-2005, 2006-2010 and 2011-2015. The value of each variable for each time period is taken to be equal to the average of all values of the specific variable for the specific time period. The selected countries are presented in Figure 4.1, while Figure 4.2 illustrates each country's GDP robustness compares to other and each country's abbreviation. Finally, the explanatory variables, in alphabetical order, are given in Table 4.1.

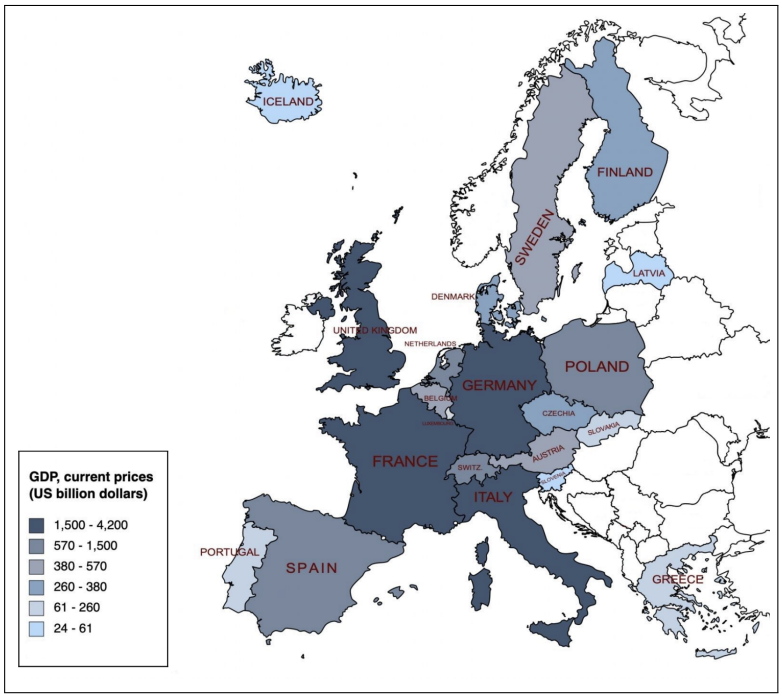


FIGURE 4.1: Mercator projection of examined countries

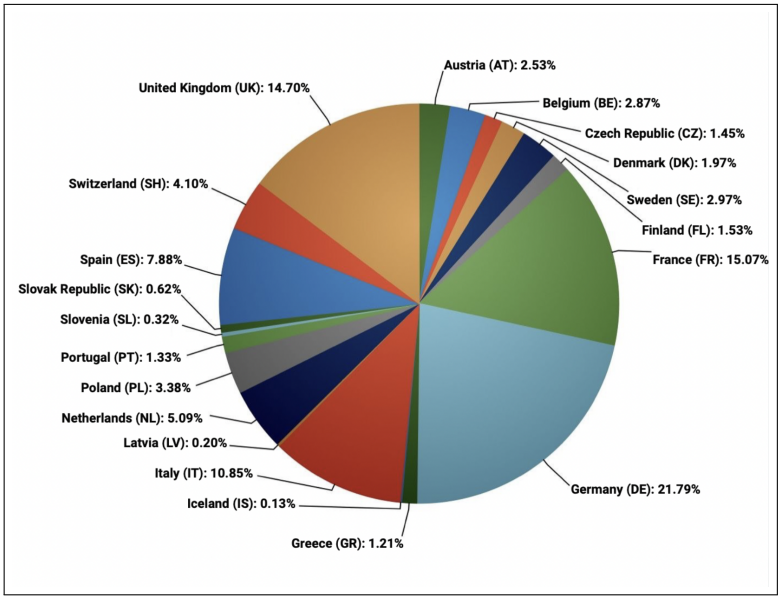


FIGURE 4.2: Percentage of a country's GDP robustness compared to the others

Compensation of Employees	Gross Domestic Product (GDP)	Median Age of Population	Total Household Savings
Consumer Price Index (CPI)	Imports of Goods & Services	Net Number of Migration Flows	Total Household Spendings
Current Account Balance (CAB)	Inflation	Net Number of Births	Total Labor Force
Demographic Dependency	Investments	Private Sector Debt	Total Saving Rate
Exports of Goods & Services	Long-term Interest Rates	Short-term Interest Rates	Unemployment Rate

TABLE 4.1: Selected possible explanatory variables

Some of the variables in Table 4.1 are directly related to Pension Expenditures like Inflation, CAB, GDP and Unemployment Rate ,while others are indirectly related. The purpose of this analysis is to identify those variables/variables that affect Expenditures. The relevant analysis is presented in Section 4.3, Section 4.4, and Section 4.5.

Note that all variables have been standardized using the average and standard deviation. The standardization ensures that all variables are measures on the same scale and as a result we avoid the possibility of recognizing falsely one or more independent variables as significant. For comparative purposes, the standardized values have been used throughout this work.

4.3 Dimension Reduction

4.3.1 Discarding Variables Technique

The 20 explanatory variables emerged from various literature studies, although some of them might not have the expected amount of impact in the formation of Pension

Expenditures as one would have thought. In order to partially "clean" the limited useful information from the data, which will result in a more efficient implementation of PCA, the Beale et al. (1967) discarding variable technique will be used as a preliminary step of the dimension reduction process. In all three time periods this technique was not completed; instead it was chosen to be stopped when, based on theoretical considerations, a very important variable was proposed for exclusion.

Based on the above discussion and taking into consideration the significance of GDP we arrived at the conclusion that, irrespectively of the time period, only those variables extracted from the original data before GDP must be excluded from further analysis, while all other variables should remain and considered for the next step of the reduction process.

As a result, the variables *Exports of Goods and Services*, *Total Household Spendings*, *Short-term Interest Rates*, *Total Household Saving*, and *Total Saving Rate* have been extracted from all datasets under examination.

The reduction process continues with the implementation of the PCA using the remaining 15 variables common to all three time periods. For research purposes, the standardized data can be accessed through this link.

4.3.2 Principal Component Analysis

In this section we apply the PCA procedure (Artemiou and Li (2009), Artemiou and Li (2013), Jolliffe (1972), Smallman et al. (2018)) as the main step of the dimension reduction procedure and obtain the full 15 principal components for each time period with the corresponding eigenvalues ranging from almost eight to nearly zero. PCA was chosen due to the multicollinearity issue, namely of the existence of high correlated covariates in the data set (correlations among more than 30 pairs of X_i's range from $|0.5|$ to $|0.98|$). The thought behind PCA is the use of an orthogonal transformation to convert a data set with interdependent variables into a new one with uncorrelated variables (principal components), which are arranged in such a way so that the first ones maintain the greater part of the variance that exists among all original variables. With this procedure the reduction of the dimension of the original data set is achieved while leaving unchanged as much as possible, the variation (Jolliffe (2002)).

Based on the overall results and the fact that we wish to avoid losing important information, we conclude that the first seven components should be kept regardless of the eigenvalues, because they retain a considerable amount of the total information/variability. The described variability played a key role in our decision since the intention was to keep that many components, so that a considerable proportion of the original variability will be described by the components chosen. It should be pointed out that the seven retained components have variability around 90% of the original variability of the data for each of the three time periods.

Note: *To determine which variables are significant in each component, the procedure used was the following. For the first two of the seven selected components we keep as significant the variables for which the absolute value of the associated coefficient is at least equal to 0.70. Although there is no specific rule, a proportion of around 0.70 is considered to be satisfactory in retaining a sufficient amount of information.*

Table 4.2 presents the most significant variables based on the components (coefficients) as a result of the PCA method, for all three datasets examined.

Dataset	2001-2005	2006-2010	2011-2015
1st Component	GDP (.96)	GDP (.96)	GDP (.98)
	Imports of Goods and Services (.94)	Imports of Goods and Services (.93)	Imports of Goods and Services (.95)
	Inflation (.89)	Inflation (.94)	Inflation (.76)
	Investments (.79)	Investments (.83)	Investments (.81)
			Net number of Migration Flows (.95)
	Number of Births (.90)	Number of Births (.90)	Number of Births (.90)
	Private Sector Debt (.93)	Private Sector Debt (.95)	Private Sector Debt (.87)
	Total Labor Force (.91)	Total Labor Force (.93)	Total Labor Force (.92)
2nd Component	Median Age of Population (-.70)	Median Age of Population (.71)*	Unemployment Rate (.81)**
	Long-term Interest Rates (.75)		
3rd Component	CAB (.70)	CAB (.65)	CAB (-.71)
4th Component	Unemployment Rate (-.80)	Unemployment Rate (-.54)	CAB (-.46)
5th Component	CPI (-.45)	Dempographic Dependency (.46)	Compensation of Employees (-.42)
6th Component	Compensation of Employees (-.52)	Compensation of Employees (-.60)	Compensation of Employees (-.46)
7th Component	Investments (-.33)	Investments (-.40)	Long-term Interest Rates (.34)

TABLE 4.2: Principal Component Analysis – The seven primary components

*The second highest variable coefficient belongs to the CPI (.68)

**The second highest variable coefficient belongs to the Median Age of Population (.65)

The first component, denoted by Z_1, in all three datasets holds at least 50% of the total variation of the dataset, while the second one, denoted by Z_2, holds roughly 20% of it. The rest of the components contain the remaining percentage of variation. The variables presented in Table 4.2 are the ones that emerge as important and play the main role in the formation of each component, without signifying that the rest should be omitted or discarded. Regarding the interpretation, the first component in all time periods can be viewed as the average of the displayed variables appearing in Table 4.2 and representing macroeconomic, demographic and microeconomic variables. On the other hand, the second component in the time periods 2001-2005 can be viewed as revealing a comparison between the Median Age of Population and the Long-term Interest Rates, while in the time periods 2006-2010 and 2011-2015 presents the average between the Median Age of Population with CPI and the Unemployment Rate with the Median Age of Population, respectively.

By construction, the first component is considered to be the most important, in which the analysis is primarily based. Having said that, we observed in the above analysis, that in all three datasets the variables that were significant in every component were almost always the same, with the variable playing the primary role and having the most influence in each of the three sets being GDP.

However, it should be pointed out that there is one important exception. Indeed, in the third time-period the Net Number of Migration Flows has been found to be significant in the first component. This variable might have an impact in the modelling process that was possibly not as important in the past as it is in this particular time period. This can be due to two very important events that have begun to emerge in Europe since 2010, the European Migrant Crisis (Garcia-Zamor (2018),

Lendaro (2013)) and Spanish, Icelandic, Portugese and Greek Economic Crisis (Gibson et al. (2014)).

Table 4.2 reveals, according to the model used, not only the importance of each variable X_i considered in the first place in assessing the value of Y, the public Pension Expenditures divided by the GDP, but also the distinction between the three periods of time examined as well. Table 4.2 also ranks the 15 variables used in each time period by each one's importance. So the interpretation should be focused actually on the next aspects: (i) The ranking of the variables according to their importance at evaluating Y in one period at a time, for every three years' time examined, (ii) The behavior of each variable X_i in respect to the different rankings between the three periods of time.

Public Pension Expenditures / GDP	Y	**Signs**	**Variable Classification**
GDP	X_1	-	Macroeconomic
Unemployment Rate	X_2	+	Macroeconomic
Total Labor Force	X_3	-	Macroeconomic
Imports of Goods and Services	X_4	-	Microeconomic
CAB (Negative/Positive Amount)	X_5	+/-	Macroeconomic
Investments	X_6	-	Macroeconomic
CPI	X_7	-	Macroeconomic
Median Age of Population	X_8	+	Demographic
Number of Births	X_9	-	Demographic
Net Number of Migrant Flows	X_{10}	-	Demographic
Demographic Dependency	X_{11}	+	Demographic
Inflation	X_{12}	-	Macroeconomic
Long-term Interest Rates	X_{13}	+	Macroeconomic
Private Sector Debt (Negative Measure)	X_{14}	-	Microeconomic
Compensation of Employees	X_{15}	-	Macroeconomic

TABLE 4.3: Correlation signs between Y and X_i based on literature studies

Considering the 15 variables affecting the cost of the pension system as random variables, an important aspect to examine is the correlation between each X_i and Y in Table 4.3. Although the sample of years and data availability might not be so sufficient as to verify 100% the theory, which for the time remains beyond the scope of this manuscript, it is useful to take Table 4.3 (For reference see Barr (2006), Blanchard (2000), Bonoli (2003), Carone et al. (2016), Diamonds (2001), Franco et al. (2006), Garcia-Zamor (2018), Holzmann (2009), Marcinkiewicz and Chybalski (2014), Muto et al. (2016), Pagès (2015), Plamondon et al. (2003), Samuelson (1958), Schneider (2005)) as an explanatory summary for variable correlation signs and classification into macroeconomic, microeconomic and demographic type of variables.

4.4 The Modelling of Pension Expenditures

In this section we proceed with the Stepwise Regression Analysis (Anderson (2009), Scheffe (1999), Sheather (2009)) using the seven components of PCA from the previous section as independent variables and Y = logit(Pension Expenditures as percentage of GDP) as the dependent variable. Our intention is to identify the significance

of each component (independent covariate) and obtain an "ideal" model for the Expenditures for descriptive as well as predictive purposes. The logit transformation was decided to be used in order to achieve the linearity between the dependent variable and each independent one as well as the homoscedasticity of the residuals.

4.4.1 Model Selection, Assessment and Comparison

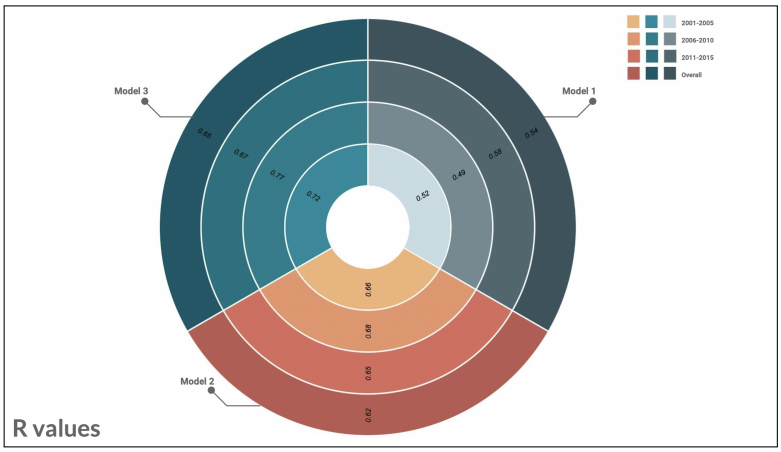


FIGURE 4.3: Resulting R values for all four examined models

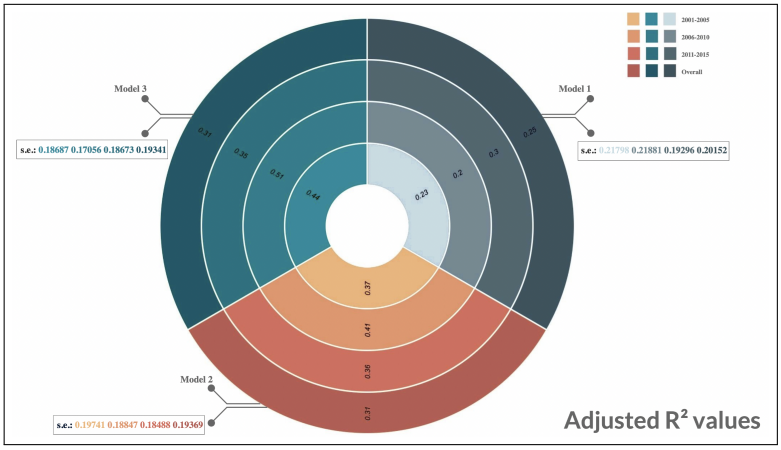


FIGURE 4.4: Resulting R^2_{adj} values for all four examined models

Figure 4.3 and Figure 4.4 contain the top 3 of 7 models, with the omitted ones being associated with at most 2% improvement.

Model 1:

$$Y = \beta_0 + \beta_1 Z_1 + \epsilon \;\; or \;\; Y = \beta_0 + \beta_2 Z_2 + \epsilon$$

Model 2:

$$Y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \epsilon$$

Model 3:

$$Y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_j Z_j + \epsilon,$$

where $j = \{4, 7, 5, 6\}$ and corresponds to {2001-2005, 2006-2010, 2011-2015, overall} datasets respectively.

The eigenvalues of the top 3 models range from 7.7 to 1.4, while the omitted ones range from less than 1 down to zero, which is in accordance with the well-known Kaiser-Guttman rule (Guttman (1954), Kaiser (1960)). Note that Figure 4.3 contains the R values for the three predominant models, while Figure 4.4 contains the corresponding R^2_{adj} values as well as the corresponding standard error (s.e.) of the model. For all three models, a different color-based categorization takes place while each time period is based on a different shade-based categorization.

Based on the results for all three datasets as well as the Overall Model, which is the average model of the three datasets combined, we conclude that the most statistically significant variables are Z_1 and Z_2.

Note that for modelling purposes both variables are used in their free form that contains not only the significant variables in Table 4.2, but all 15 variables resulted after the implementation of the Beale et al. (1967) technique.

Note that in two instances a third variable appeared to be of some significance. Variable Z_4 (for 2001-2005) and variable Z_7 (for 2006-2010) appear to have some contribution but we choose not to include them in the analysis not only for homogeneity purposes but also due to the fact that Z_1 and Z_2 according to PCA, contain more than 65% of the total variation while Z_4 and Z_7 explain a small (statistically not significant) amount of the total variation. Hence, we proceed below with the Multivariate Analysis of the dependent variable Y with Z_1 and Z_2 as the only independent ones for all 4 models under investigation. Note that (i) All results were interpreted with $\alpha = 5\%$ and (ii) For the implementation of the regression analysis, the usual assumptions of independence, normality and homoscedasticity of the residuals must be fulfilled as well as the linear relationship between Y and Z_i, $i = 1, 2$. In addition, multicollinearity should be verified.

The appropriateness of PCA applied in the datasets under investigation has been verified by the validity of the assumptions associated with PCA (O'Rourke et al. (2005)) including the linearity ensured by the transformation considered in the above analysis. Furthermore, note that the selection of the first two components ensures that the amount of variability explained is sufficiently high to retain a considerable degree of the internal structure of the datasets.

4.4.2 Regression

Table 4.4 and Figure 4.5 provide the regression analysis results for Y with respect to Z_1 and Z_2 defined in the previous Section.

	2001-2005			**2006-2010**			**2011-2015**			**Overall Model**		
Model	SS	F	Sig.	SS	F	Sig.	SS	F	Sig.	SS	F	Sig.
Regression	.522	6.699	.007	.540	7.606	.004	.439	6.415	.008	.402	5.354	.016
Residuals	.663			.604			.581			.638		
Total	1.185			1.144			1.020			1.039		

TABLE 4.4: Analysis of Variance for all datasets based on the selected model

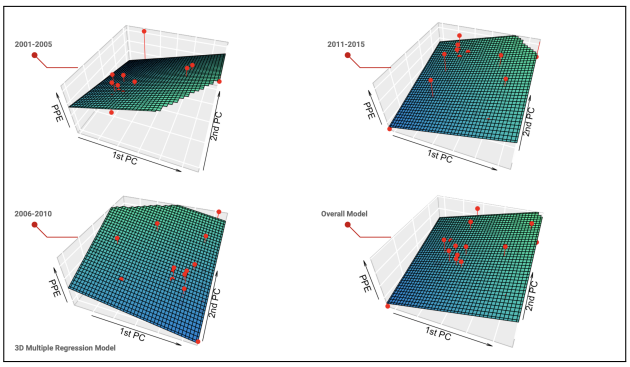


FIGURE 4.5: Regression Coefficients of Pension Expenditures

Figure 4.5 display the multiple regression models ($Y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2$) for all examined datasets, in a 3D space.

Based on the results in Figure 4.3 - Figure 4.5, and Table 4.4, one could make the following observations:

1 The R coefficient ranges from 49% to 77% and the Adjusted R^2 from 20% to 44% for all 4 models.

2 From the F-test of analysis of variance (ANOVA) table we conclude that there is at least one independent variable which is statistically significant (p-values range from 0.004 to 0.016), which is verified by the appropriate t-test which states that at least one of Z_1 and Z_2 is statistically significant for each of the 4 models (p-values range from 0.005 to 0.133).

3 The assumptions of Independence and Homoscedasticity between the residuals as well as the linearity are fulfilled for all 4 models. Furthermore, the collinearity is small in all cases, while residual analysis reveals that there is a deviation from Normality in all cases.

4.5 Macroactuarial Justification

In this Section we compare the observed values of the dependent variable Y with the estimated values that have been obtained from the regression for each of the four datasets. Note that in Figure 4.6 and Figure 4.7 the observed values are represented with a solid line (–) while the expected values with a dashed line (...).

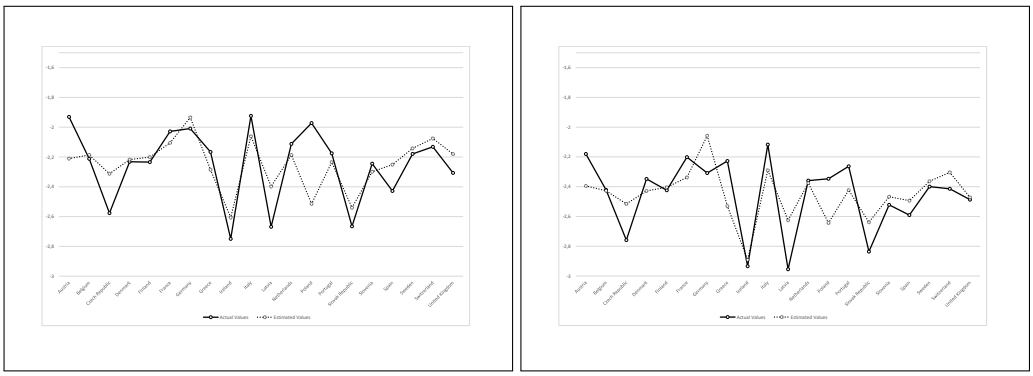


FIGURE 4.6: 2001-2005 and 2006-2010 model

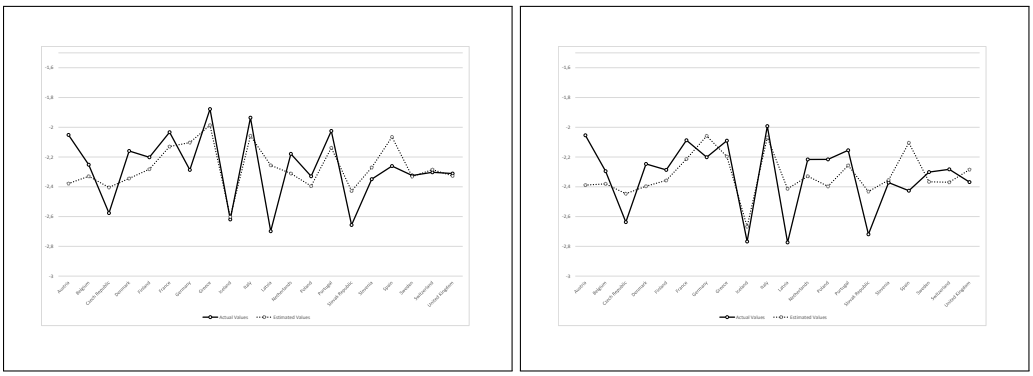


FIGURE 4.7: 2011-2015 and Overall model

Summarizing, based on Figure 4.6 and Figure 4.7 we can state that our model selection fits well for most countries in all three datasets as well as in the Overall Model.

According to the analysis, the most divergent countries are AT, CZ, LV, PL, SK and ES. We can observe that 4 of the total 6 countries are part of Eastern Europe as it has been registered by the United Nations. An explanation of this fact is provided by Müller (2001) who states that *"The retirement schemes in several East European countries underwent fundamental change in recent years"*. The common thing between all these countries is that they have faced significant migration reversal of trends between the three periods (see Figure 4.8) large enough in comparison to their overall population (Eurostat (2019)). In fact, AT, as so classified as a western European Union (EU) country, was always a receiving migrant country but flows have grown excessively from 2009 and onwards, creating an unbalanced situation under formation. Moreover, during the 2011-2015 period examined when X_{10}: Net Number of Migration Flows variable is appearing very important, a kind of opposite effect is being observed for the rest of the countries (ES, CZ, LV, PL and SK) four of which are classified as Eastern EU countries. These countries, except CZ and SK have been generally sending migrants to the rest of the EU. After 2009, they experienced a disproportionate reversal of the trend (ES, CZ and SK) or a remarkable enhancement of the trend of sending migrants (LV and PL) as a relatively large proportion of their population. Also, the GINI index[2] of all these countries except AT is high enough, classifying CZ, LV, PL, SK and ES together as countries with remarkable income inequalities connected with increased $Y's$ (Marinescu and Manafi (2017)). In fact, the heterogeneity of these countries compared to the observed other countries under investigation, might have an impact in the creation of deviations.

[2]The Gini index, also known as the Gini coefficient or ratio, is a statistical dispersion measure intended to represent income or wealth inequality within a nation group.

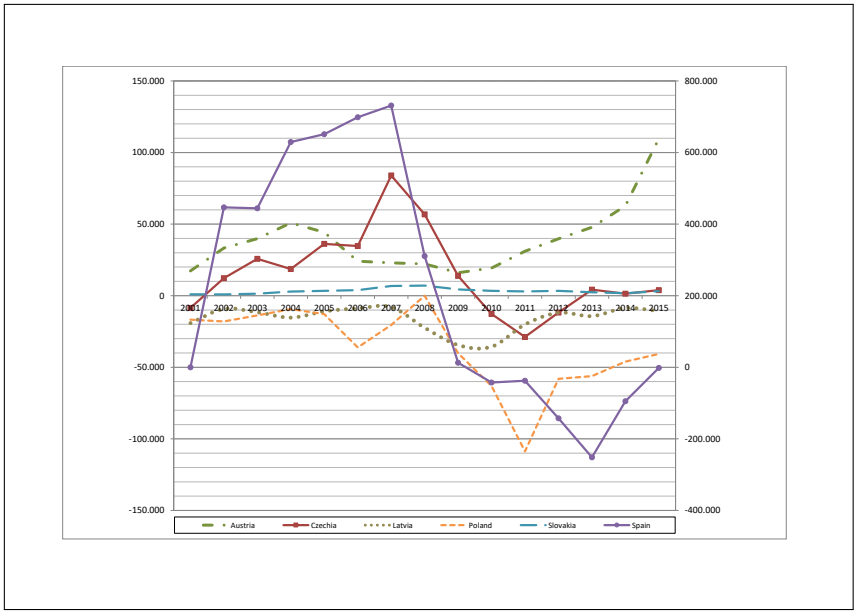


FIGURE 4.8: Net Number of Migration Flows 2001-2015 for AT, CZ, LV, PL, SL (left axis) and ES (right axis)

4.5.1 Macroactuarial Interpretation

As indicated earlier, Table 4.2 reveals both the importance of each $X_i's$ in assessing Y and the distinction between the 3 time periods. Also, in all three periods, period 1 (2001-2005), period 2 (2006-2010) and period 3 (2011-2015), seven variables emerge as the most important in the first component, X_1: GDP, X_3: Total Labor Force, X_4: Imports of Goods and Services, X_6: Investments, X_9: Number of Births, X_{12}: Inflation and X_{14}: Private Sector Debt.

Table 4.3 revealed the correlation signs between Y and X_i based on literature studies. From the statistical analysis results we almost came to the same conclusions about correlation, apart from X_5 and X_8, as it can be seen in Figure 4.9. The reason for this, apart from the limited periods examined, may be that X_5: Current Account Balance, for more than half of the countries is negative in all three periods examined. Germany is always having a positive CAB with a negative correlation with respect to Y, which due to the magnitude, increases disproportionally from all country averages. Also, the average correlation of X_8: Median Age of Population, although negative appears to be low (-9%).

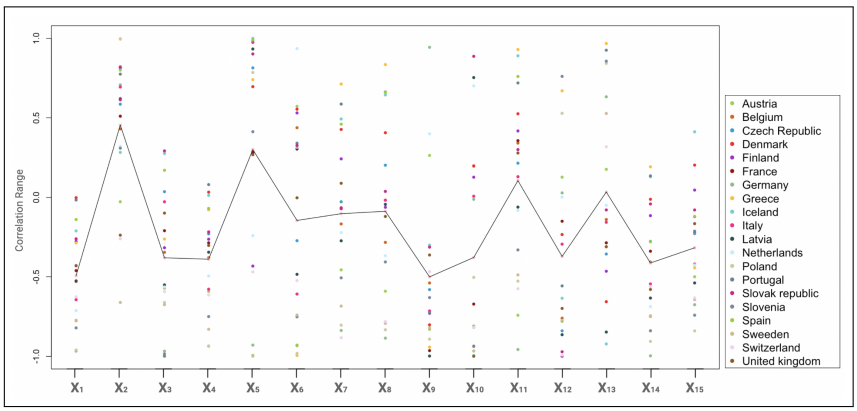


FIGURE 4.9: Correlation plot of Y with all X_i's

In Figure 4.9 we can see the correlation between PPE and each X_j for all countries. Each country is classified by a unique colour. The black line connects the average correlation points, i.e., the average correlation of Y with the X_j for all the countries.

As mentioned above, the first two components and by extension the two new variables represent the core in evaluating Y. Focusing solely at the variables in Table 4.2 that contribute significantly at each PCA component, and Figure 4.9 and Figure 4.10, we reached the following:

A. Intra period interpretation

i. Period 1: As mentioned above, seven variables emerge as important in the first Component. The ranking of importance of those variables is by order of maximum to minimum importance X_1, X_4, X_{14}, X_3, X_9, X_{12} and X_6. The four most important macroeconomic variables are X_1, X_3, X_9 and X_{12} and the two most important microeconomic variables are X_4 and X_{14}. Actually, the net indirect taxes upon X_4: Imports of Goods and Services (as Value-added tax, VAT) represent one of the three components of X_1: GDP from the income side (Blanchard (2000)). Another component of X_1 from the same side is the X_{15}: Compensation of Employees (Blanchard (2000)) representing from 20% up to 80% of the GDP. Results for this period imply that almost all information needed for Y is carried by X_1, since X_{15} belongs to the 6th Component. Also X_{14}: Private Sector Debt is mostly affected by public pension costs because contributions to social security represent an additional tax which may asymmetrically burden the private sector's economic activity. X_{12}: Inflation is usually the same measure as the GDP deflator and represents the part of the GDP's evolution due to the increase of the general level of prices produced in the countries examined, irrespective of real growth. So when Y is rising, Unemployment rises and X_{12} falls. Vice versa when Y is falling, distortions in income distribution in the economy fall and real pensions are indirectly increased when lower Inflation enables pensioners to buy the same or even more goods with their decreased pensions.

The demographic variable X_9: Number of Births, according to the model, is the most important demographic variable because births accumulate young population yearly and immediately affect X_{11}: Demographic Dependency as well as X_8: Median Age of Population. PAYG (*Pay As You Go* is the practice of financing expenditures directly with contributions of social security rather than investments) systems, which have prevailed in the last decades and bear almost 100% of pension costs, if properly function as pension redistributive systems directly from workers to the pensioners, are most sensitive to the population structure. Three to five workers are needed to support everyone pensioner.

X_{13}: Long-term Interest Rates belong along with X_8 to the 2nd Component. Both these variables are affected by the country's deficit or surplus accumulation over time (Diamonds (2001)). Lower $Y's$ leads to a more favourable interest rate in the long run for a country to pump money from the capital markets.

X_5: Current Account Balance belongs to the 3rd Component; it represents yearly the output Gap of a country's economic activity. It is a deficit or a surplus. If Y falls, the amount of deficit may fall, or a surplus might be more feasible (Diamonds (2001)). Actually, Y is negatively correlated with surplus and positive correlated with deficits.

X_2: Unemployment Rate belongs to the 4th Component, it is the fraction of the number of people being able to work but have lost their jobs (unemployed) to the

number of people who work plus the unemployed (the Total Labor Force). The number of unemployed people is also a component of X_3: Total Labor Force; Much of its information is included in X_3.

From the results X_7: CPI, representing the price increase of the goods consumed in the countries, bears minor importance belonging to the 5th Component. This shows that during the short run of this period CPI adds limited information in comparison to the information carried by X_{12}: Inflation, i.e., the total overpricing of all output of the economy. In the long run, these two indicators will coincide. The demographic Variables X_{10}: Migration Flows, and X_{11}: Demographic Dependency have been completely eliminated from Period 1 by the method of PCA so they do not contribute at all to the determination of Y. This might be attributed to the fact that Migration Flows might have not reached a point of affecting the population structure with the demographic variables X_9, as well as X_8, carrying all the information needed to assess Y in Period 1.

ii. Period 2: The same seven variables as in Period 1 emerge in this time-based dataset as significant in the 1st Component but with a different order of importance: X_1, X_{14}, X_{12}, X_3, X_4, X_9, X_6.

In this period only X_8: Median Age of Population belongs to the 2nd Component. X_{13}: Long-term Interest Rates has been eliminated completely meaning that states are performing reforms (Carone et al. (2016)) trying to pass a part of the PAYG system to funded systems not affecting so much the rates of publicly issued bonds. Also, X_7: CPI has been exiled meaning that it may have grown almost identical to X_{12}: Inflation. X_{10} remains unimportant according to the modelling also for this period of time.

X_5: Current Account Balance and X_2: Unemployment Rate, X_{15}: Compensation of Employees and X_6: Investments still belong to the 3rd, 4th, 6th, and 7th Component respectively showing a kind of steady state in the way of affecting Y.

However, X_{11}: Demographic Dependency has earned an advanced place of importance in the 5th Component, entering as a third important demographic variable. As time goes by, the aging of the population advances and demography plays a more and more crucial role concerning public pension costs.

iii. Period 3: Again, the same seven variables emerge in the 1st Component, but an eighth variable, X_{10}: Migration Flows has entered into the 1st Component for the first time –although completely meaningless in the two previous periods examined. The new order of importance is X_1, X_{10}, X_4, X_3, X_9, X_{14}, X_6 and X_{12}. Also, three demographic variables describe Y for this period, in a different way, with X_9, X_{10}, and X_{11} bearing all the demographic information. However X_2: Unemployment Rate is more important belonging to the 2nd Component, outlying the growing dependency of the pension system from the Labor Market described by two key variables X_3: Total Labor Force and X_2 itself. Also X_5 the CAB has gained two important places instead of one during the two previous periods composing both the 3rd and the 4th Component. Again, here someone can observe the growing influence the pension system has on the public budget. In this period X_8: Median Age of Population has been eliminated together with X_7: CPI.

B. Inter period interpretation

As expected from the three periods examined, the key variables, mainly or primarily affecting the Pension Expenditures are the variables X_1, X_{10}, X_4, X_3, X_9, X_{14},

X_6, X_{12}. Reforms performed by countries reorient the importance of the 15 variables considered, pushing to lower dependency for the pension costs from the public budget. However, demography gains an advancing role as time goes by in assessing Y. Migration also, since it alters the population structure of the countries studied, will be growing more and more important. Macroeconomic variables remain strongly interrelated with Y and states should take Expenditures of much consideration in the economic cycle when designing reforms.

4.5.2 The Migration Effect

From experience and the literature concerning Y, the Pension Expenditures (Barr (2006), Blanchard (2000), Holzmann (2009), Marinescu and Manafi (2017), Müller (2001), Muto et al. (2016), Pagès (2015), Plamondon (2003), Schneider (2005)), have been observed to bear correlations between it and various demographic, macro, or micro variables mentioned in this work. These correlations specify a special mixture of features that characterize each country examined. So, for countries which traditionally accommodate migrants, having developed in the meantime, work and educational inclusion policies, migration is negatively correlated with Y because it drops the median age of the population and favours the reduction of Expenditures.

In this work, variable X_{10} represents Migration Flows, i.e., immigration minus emigration population movements between the countries examined and third countries outside them. The difference in Migration Flows with the other measures (or variables) examined –except births– is that it represents population changes and not absolute numbers. In fact, these flows accumulate more people in the population of European countries examined apart from Latvia and Poland, where they show a continual negative trend. So on average between 2001 and 2015 the accumulation of migrant population has risen to more than 3% of the total population of the countries, and amounts to almost 18 million people. There is some general evident that migration is negatively correlated with Expenditures (Marinescu and Manafi (2017)). From data sources used (Eurostat (2019)), countries bearing big correlation of Y and X_{10} have low correlation of Y and X_{11}: Demographic Dependency ratio and vice versa, see Alluvial Diagram of Figure 4.10.

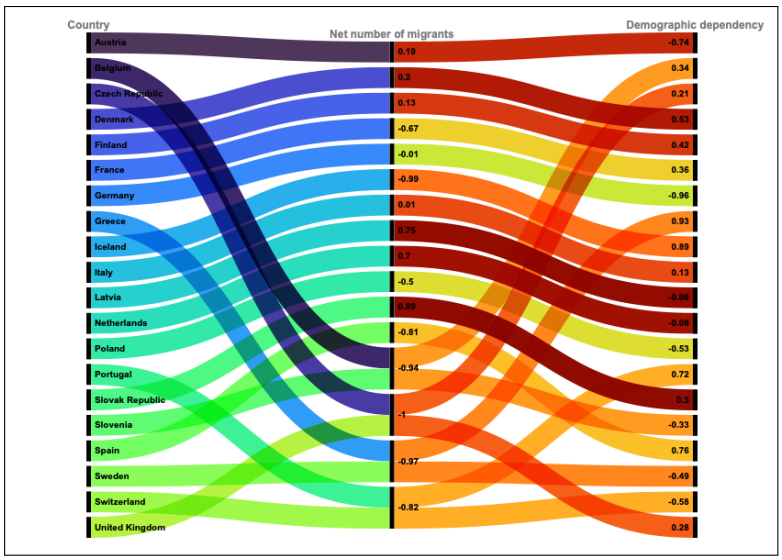


FIGURE 4.10: Alluvial Diagram of per country PPE correlation with Migration Flows and Demographic Dependency

From Figure 4.10 it is evident that 12 out 20 countries (BE, CZ, FR, GR, IS, PL, PT, SL, ES, SE, CH, GB) present negative Y and X_{10} correlations, giving a support to theory. For half of them (BE, CZ, GR, IS, PT, GB) the negative correlation of Y and X_{10} is combined with the high correlation of Y and X_{11}. This may be attributed to the fact that either they do not incorporate migrants into the country's economy, or their pension system is almost independent of demographic dependency and it is based on minimum flat rate pensions, as in the GB. On the other hand, for 4 countries (AT, DE, LV, NL) the opposite effect is shown (Positive Y and X_{10} and negative X_{11} correlation). For Latvia the low demographic dependency correlation might correspond more to the fact that Migrant Flows appear to be negative for all five years' periods monitored and also to its vast pension reforms even before 2001 reducing the pension replacement ratio[1] to 32%, the benefit ratio[2] below 30% and Y around 8%, the lowest in the EU countries (Carone et al. (2016)).

So, for all countries, migration plays a significant role since the EU countries continue to attract large immigration flows, which grows more important as far as this trend continues. In this context, a future analysis concerning the 2016-2020 period will bring again the migration as the 1st component of importance in measuring Pension Expenditures.

4.6 Conclusions

In conclusion, in this work, we suggest the same model for all three five-year period datasets as well as for the Overall time period one based on PCA and regression analysis for the modelling of the PPE of European countries since 2001. This model consists of the first two components out of a total of 15, which contain more than 70% of the total information and variability of the original data.

The idea behind this work was to create a model for a plethora of countries, at first within Europe and later a worldwide one, so we can compare them and also to achieve the following tasks. Firstly, by reducing the dimensionality of the original dataset, we obtain a more "easy to use and handle" model and can apply various statistical methods and techniques without losing the accuracy and information of the original variables. Secondly, we are able to limit or eliminate the existing multicollinearity, and therefore achieve a more accurate model interpretation of the PPE.

The model developed provides, with a minimum average error of fewer than 6‰ for each time period, accurate results for the PPE. Using this as a first step, it is possible, depending on the data available, to develop in the immediate future, an evolved time series model that would be capable of predicting the Expenditures for 10-15 future years from the base year. The forecasting model could be used by any state that wishes to predict future Pension Expenditures based on its economy. This calculation primarily serves either as an estimate by itself or as a confirmation technique for the calculation of Expenditures made by other means.

[1] Ratio of the last salary to the first pension amount

[2] Ratio of the country's average pension to the average salary

Chapter 5

Feature Selection Partial Least Squares (FS-PLS): The Utilization of Partial Least Squares for Simultaneous Feature Selection and Extraction.[1]

5.1 The PLS algorithm for dimension reduction

So, as stated before, PLS is considered to be an effective dimension reduction technique when it comes to obtaining an optimal statistical model. However, like many similar feature extraction techniques, we end up with a model that involves all original variables, significant, or not. What if there was a way to take advantage of PLS algorithm in order to utilize it as a variable selection technique? The FS-PLS is a novel approach that allows the researcher to use the PLS procedure to remove non-significant variables from the original dataset and obtain a statistically significant model with minimum dimension when PLS is applied. FS-PLS provides a new dataset with simpler structure than the original one and still when its implementation is compared to PLS and PCA, the model arises from FS-PLS is more efficient than the corresponding models of PLS and PCA. This "superiority" is due to the fact that the constructed model of FS-PLS is easier to interpret since all irrelevant variables have been removed.

The beta coefficients (β) that emerge from the PLSR in conjunction with the number of selected latent variables can be seen as a general rule of thumb for disregarding variables from a dataset. We propose the following rule to determine if a variable is significant:

Let us assume a model with X_j, $j = 1, \ldots, m$ independent variables and let v be the number of latent variables that have been selected as optimal from the PLS regression of the aforementioned model. Let us also assume that β_j^v being the corresponding coefficient of X_j variable in the v-latent variable (each latent contains all original variables). Now, let us define Equation 5.1 as follows:

$$|\beta_j^v| \leq c, \tag{5.1}$$

[1]The results of this Chapter have been submitted for publication as:
Beki, E., Karagrigoriou, A., and Ntotsis K.: The Utilization of Partial Least Squares for Simultaneous Feature Selection and Extraction (2022)

where $c \in [0.05, |max\{\beta_j^v\}/2|)$ is a pre-determined non-negative value close to zero and $|max\{\beta_j^v\}|$ is the maximum (absolute) value that exists in the coefficient matrix of the selected v latent variables. If Equation 5.1 is satisfied for the j-th variable, i.e. $|\beta_j^1| \leq c$, and $|\beta_j^2| \leq c$, ... and $|\beta_j^v| \leq c$, then this variable can be labelled as non-significant. By integrating this β-based constraint in the PLS regression, it is feasible to discard the insignificant variables and still maintain a robust model. A fixed value c is expected to complement effectively all other aspects (purpose of the study, researcher's judgement, etc.) of the decision-making process. In that sense, it can be considered as a rule of thumb and is in the judgement of the researcher which value of c is the one that results the optimal PLS model without underfitting or overfitting the model under consideration. We recommend a step procedure of 0.05 units (i.e. 0.05, 0.10, 0.15, etc.) until model underfitting is observed based on the model selection criteria.

The FS-PLS algorithm consists of a two level implementation of the PLSR algorithm. Initially, PLS method is applied on the original dataset and the regression coefficients of the original variables are estimated with the use of models consisted of up to three latent variables. Those variables with absolute values of regression coefficients lower than the testing threshold in all three models are considered insignificant for the prediction of response variable and they are removed from the dataset. This distinction between the variables is followed by the application of Partial Least Squares Regression to generate predictive models. Their competency is evaluated based on information criteria, such as AIC, Adjusted R^2 (R^2_{adj}), RMSECV and Adjusted Wold's R criterion, that lead to the final model selection.

The following algorithm displays the proposed procedure

Algorithm 4 Pseudocode for FS-PLS

Input: A data set consisted by a $n \times p$ matrix X and a $n \times 1$ matrix Y, where each X_j and Y column represents a variable, and a constant threshold c.
Output: A data set consisted of the minimum variables that can result in the optimal PLS model.

Step 1: Application of PLSR on original data for the evaluation of regression coefficients.

Step 2: Usage of model selection criteria for number of optimal latent variables determination

Step 3: Application of the constrain proposed in Equation 5.1 for the location of the statistically insignificant variables.

Step 4: Removal from the input dataset the variables that **Step 3** indicate as insignificant

Step 5: Repetition of **Step 1** on the minimized original data

5.2 Numerical applications

In this section the application of the FS-PLS on near infrared (NIR) spectroscopy data is presented.

5.2.1 Univariate FS-PLS regression – FS-PLSR

In the first case, in the *gasoline* dataset, which is found in the `pls` package, X matrix includes 401 diffuse reflectance measurements and Y matrix is consisted of one response variable, that corresponds to the number of octanes of the total 60 observations. Due to the multicollinearity and the rate of available observations to X-variables, dimensionality reduction is demanded in order to generate a linear regression model. Applying the FS-PLS optimization, we first computed the estimators of PLS-regression coefficients of all 401 variables in models built with up to three components. Their absolute values were then compared with predefined constant c of 0.10, 0.20, 0.25 and 0.30. The final X data matrices contextually included only the predictive variables with absolute values of PLS-coefficients higher than the testing threshold in one-, two- and three-component models (1 LV, 2 LV, and 3 LV). In the next step we reapplied the PLSR method to the selected variables and the resulted models were evaluated based on AIC, R^2_{adj}, $WR^{0.90}_{adj}$, and $WR^{0.95}_{adj}$ and RMSECV. Table 5.1 and Table 5.2 summarize the results:

					1 LV		2 LV		3 LV		4 LV		5 LV		6 LV	
c	Attributes	$WR^{0.90}_{adj}$	$WR^{0.95}_{adj}$	RMSECV	AIC	R^2_{adj}	AIC	R^2_{adj}	AIC	R^2_{adj}	AIC	R^2_{adj}	AIC	R^2_{adj}	AIC	R^2_{adj}
-	401	4	4	7	203	30%	52	94%	3	97%	-2	97%	-25	98%	-36	98%
0.10	374	4	4	7	203	30%	52	94%	4	97%	-3	97%	-25	98%	-36	98%
0.20	307	4	4	7	203	31%	52	94%	6	97%	-3	97%	-24	98%	-35	98%
0.25	245	4	6	6	202	31%	51	94%	9	97%	-4	98%	-23	98%	-35	98%
0.30	217	4	4	6	202	31%	51	94%	12	97%	-5	98%	-22	98%	-35	98%

TABLE 5.1: Information criteria values of FS-PLSR models, where Attributes is the number of original variables.

In Table 5.1, the reduction in AIC values in all two-component models and the simultaneous increase of their R^2_{adj} values is noteworthy. These changes strongly indicate the outstanding enhancement of the corresponding models when the second component is retained in the model. Further, the most sufficient FS-PLSR model is proposed, the four-component model, which is based on the 0.30 testing threshold and it includes only 217 variables in X matrix, which consist of 46% of the initial observations. This choice is established in accordance with the adjusted Wold criterion, which is specialized to evaluate PLS models, complemented by the high R^2_{adj} value and the significant reduction in AIC value. It should be noted that AIC values tend to decrease as more components are added to the model. However, the rate of decrease is approximately fixed after the addition of the fourth component. Moreover, the criteria values of the models that resulted from the thresholds 0.25 and 0.30 are alike, though the latter constraint conveys to further dimensional reduction. At this point, it should be mentioned that more restrictive thresholds were tested; they were found to lead to over-fitted models and rejected.

The results of the PCA regression (PCR) models, generated with the datasets arising from the aforementioned thresholds, are displayed in Table 5.2. As the most adequate model is proposed the five-component model of the last threshold, since R^2_{adj} value is close to 1 and AIC value does not change sufficiently with the addition

of more components in the model. In contradiction to the FS-PLSR models, the inclusion of the second component does not improve the model performance in any case, while the minimization of RMSECV values proposes much more complicated models than in FS-PLSR cases.

		1 PC		2 PC		3 PC		4 PC		5 PC		6 PC	
c	RMSECV	AIC	R^2_{adj}	AIC	R^2_{adj}	AIC	R^2_{adj}	AIC	R^2_{adj}	AIC	R^2_{adj}	AIC	R^2_{adj}
-	17	213	17%	215	17%	192	43%	6	97%	6	97%	8	97%
0.10	17	213	17%	215	17%	189	46%	6	97%	6	97%	7	97%
0.20	15	213	17%	214	17%	174	58%	6	97%	5	97%	7	97%
0.25	15	213	17%	214	18%	147	73%	7	97%	5	97%	5	97%
0.30	14	213	17%	214	18%	133	79%	7	97%	5	97%	5	97%

TABLE 5.2: Information criteria values of PCR models

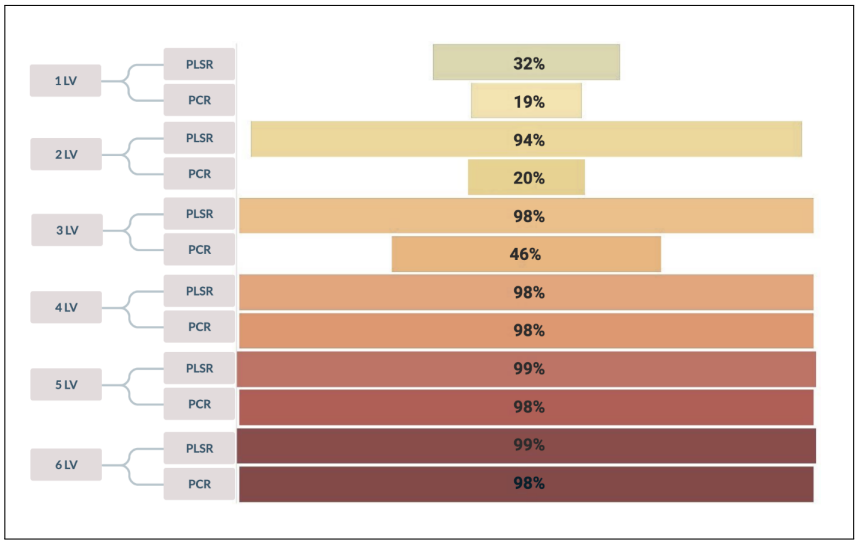


FIGURE 5.1: Percentage of explained variability of FS-PLSR and PCR models

Additionally, taking into account the R^2_{adj} criterion and the percentages of explained variability in the models, as displayed in Figure 5.1, we conclude that in FS-PLSR the two-component and three-component models can lead to reliable results, preserving the advantage of visualization. Note that all c constrains resulted in similar explained variability and thus only one's results are being presented in Figure 5.1. These FS-PLSR models expose high R^2_{adj} values, while they leave unexplained a negligible percentage of the response variable. In PCR instead, the inclusion of the first four components fails to provide a model with sufficient performance. Finally, the comparison of these methods in terms of AIC values verifies the predominance of FS-PLSR against PCR: all AIC values in PCR models (with up to three components) are significantly smaller than the corresponding FS-PLSR model (Table 5.2).

5.2.2 Multivariate FS-PLS regression – FS-MPLSR

In the second case, the performance of the proposed FS-MPLSR optimization over a multivariate response is investigated (the multivariate case of FS-PLSR will be addressed by FS-MPLSR in the remaining article). In the Corn dataset (2022), that we processed, the Y matrix consists of four variables, -moisture, oil, protein and starch,

and X matrix includes 700 NIR spectroscopic attributes. Note that the FS-MPLSR algorithmic procedure is similar to the FS-PLSR with the only deference to be the number of response variables that form the latent variables. In the multivariate case, the modelling process aims to reveal and enable chemists to predict the moisture, oil, protein and starch content in different samples. In this situation, the implementation of Ordinary Least Squares as a linear regression method would be an inappropriate choice, since the X matrix is characterized by the existence of multicollinearity. Its mitigation is achieved through dimensional reduction, based on the absolute values of the FS-PLS-regression coefficients, in a similar way as in the univariate case. The FS-MPLSR algorithm was applied on the initial dataset to estimate these values. The computation of Adjusted Wold's R criterion $WR^{0.90}_{adj}$ led to the conclusion that the sufficient modelling of the four Y-variables requires the inclusion of first 5, 21, 7 and 8 FS-PLS components respectively. Based on this conclusion and the use of testing thresholds we defined the final reduced set of predictors as the intersection of the following four subsets:

- The first subset included the variables considered as statistically significant for Y1. The absolute values of regression coefficients of these variables are higher than the tested thresholds in one- to five-component models.
- The second subset included the variables considered as statistically significant for Y2. The absolute values of regression coefficients of these variables are higher than the tested thresholds in one- to twenty one-component models.
- The third subset included the variables considered as statistically significant for Y3. The absolute values of regression coefficients of these variables are higher than the tested thresholds in one- to seven-component models.
- The fourth subset included the variables considered as statistically significant for Y4. The absolute values of regression coefficients of these variables are higher than the tested thresholds in one- to eight-component models.

The thresholds that we tested were 2, 2.25, 2.50 and they resulted in the removal of 44, 69, and 99 variables from the original dataset, correspondingly. The new reduced data matrices were then processed via the FS-MPLSR and PCR methods. Based on the values of the aforementioned model selection criteria we inferred that the third threshold examined (2.50) generated the most efficient models. Table 5.3 and Table 5.4 summarize the values. The other options led to over-fitted or under-fitted models.

	1 LV		2 LV		5 LV		7 LV		8 LV		omitted LV	21 LV	
	AIC	R^2_{adj}	AIC	R^2_{adj}	AIC	R^2_{adj}	AIC	R^2_{adj}	AIC	R^2_{adj}	$\cdots$	AIC	R^2_{adj}
Y1	1	51%	2	52%	-50	80%	-142	96%	-166	97%	$\cdots$	-366	99%
Y2	-62	27%	-62	27%	-68	37%	-87	55%	-154	86%	$\cdots$	-214	95%
Y3	76	17%	65	32%	-35	88%	-56	92%	-60	92%	$\cdots$	-181	99%
Y4	153	0%	145	13%	63	78%	28	88%	7	92%	$\cdots$	-131	99%

TABLE 5.3: Information criteria values of the FS-MPLSR model based on the remaining 601 attributes (700-99).

	1 PC		2 PC		5 PC		7 PC		8 PC		omitted PC	21 PC	
	AIC	R^2_{adj}	AIC	R^2_{adj}	AIC	R^2_{adj}	AIC	R^2_{adj}	AIC	R^2_{adj}	···	AIC	R^2_{adj}
Y1	1	52%	1	53%	-115	93%	-153	96%	-151	96%	···	-371	99%
Y2	-62	27%	-61	26%	-69	38%	-116	72%	-124	76%	···	-158	88%
Y3	76	17%	74	22%	22	68%	-24	86%	-56	92%	···	-112	97%
Y4	153	0%	151	0%	120	44%	76	74%	52	83%	···	-47	97%

TABLE 5.4: Information criteria values of the PCR model based on the remaining 601 attributes.

The optimum FS-MPLSR model retained twenty one components. The model complexity was determined in accordance with the theory (Wold et al. 2001), which states that the FS-MPLSR model should include every component that is found to be significant for at least one variable of the set of responses. In this way information in Y matrix is significantly explained, as the percentages of the explained variability are 99.95% for Y1, 97.06% for Y2, 99.43% for Y3 and 99.55% for Y4, while the overall information of the new X matrix is utilized. We can infer that the substantial dimensionality reduction that we achieved through the PLS-optimization, resulted in the generation of a unique model capable to predict the four responses at the same time, with the cost of an insignificant percentage of unexplained information. Nevertheless, a less strict consideration of the theoretical frame would yield an eight-component model with the profit of further dimensionality reduction and with the cost of a less accurate, but yet sufficient, prediction of Y2 response variable.

On the contrary, the PCR method generated four individual models, one for the prediction of each response variable, that needed twenty one components to capture 99.95%, 92.51%, 98.21% and 98.17% of the variability of the responses. These percentages, in combination with the results of the information criteria presented in Table 5.3, demonstrate that FS-MPLSR model is more adequate in all four responses.

5.3 Concluding Remarks

The aim of this study is to introduce PLS as a method for variable selection in a variety of fields, including time series analysis. Although this method is commonly used in a regression analysis, it can also be implemented in various other applications such as discriminant analysis, and hierarchical modelling. It can handle complex data sets and situations that cannot be solved by standard methods.

FS-PLS is considered optimal in evaluating more complex structures with a more realistic and holistic view. It has been proved to be a non-time consuming process and statistically efficient method with high prediction accuracy. As a recently found technique in the field, many aspects of its underlying mechanism have recently been revealed and yet, there is no strictly defined frame for its application. As a result, the method is considered to be very flexible and many modifications and experimentations can be tested. In this work the utilization of PLS approach was used as a variable selection criterion and by expansion as a dimension reduction technique. The FS-PLS procedure was able to remove up to 45% and 14% of the original variables in two frequently used datasets in chemometrics, one univariate set and one more complex multivariate one.

Although PLS is considered to be useful in small datasets, through the FS-PLS methodology it has been found to be useful in high-dimensional and/or big data analysis. Although, the applications is chosen from the field of chemometrics, the

applicability was quite wide covering biology, physics, chemistry, business, and social sciences among others.

In the univariate case, the final selected model is based on only 217 predictors out of an initial set of 401. The three-component model, which is suggested as optimum, explains the major part of information captured in the data, while it is parsimonious, with high prediction ability and can easily be used for visualizations. The comparison with the corresponding PCR model, which was based on information criteria AIC, R^2_{adj}, and RMSECV, demonstrates that FS-PLSR model gave more sufficient results.

In the multivariate case, the problem appears to be more complicated. Initially FS-MPLSR was implemented on the data out of necessity, due to the fact that correlations were observed between the response variables. We estimated regression coefficients and we determined the significant components for each response variable. We compared the absolute values of the coefficients in significant components with thresholds and then, we defined four sets of predictors, which contained the important predictors for the individual responses, respectively. Their intersection consisted the final set of predictors for the multivariate regression model. This way, in the final selected model 99 less predictors than in the initial set were included. The simultaneous process of the response variables generated a single regression model with AIC values lower than the individual PCR models in all four response variables. The increased number of constructed models in the PCR method is associated with high complexity and computational cost of the whole analysis. This, in combination with the fact that less variability is explained in the second response variable with the PCR method, leads to the suggestion that a FS-PLSR model is optimum also in the multivariate case.

Concerning possible future expansion of this work, an ENR in the FS-PLS method is under development. We aim to investigate the cooperative effects of these two techniques on high-dimensional multicollinear data in order to make a projection on a low-dimensional space and thus to construct less complex and more interpretable linear regression models of high predictive accuracy with a penalized set of predictors.

Chapter 6

Interdependency Pattern Recognition in Econometrics: A Penalized Regularization Antidote[1]

Several partially robust criteria and indices for multicollinearity have been proposed over the years, which are based either on the coefficient of determination and similar measures or in the eigenvalue-eigenvector analysis. Theil's indicator (1971), Klein's rule (1962), TOL, and VIF (Gujarati and Porter (2008)) fall into the first category while the Farrar-Glauber test (1967), the sum of reciprocal eigenvalues, Red indicator (Kovács et al. (2005)), CI (Belsley (1991), Hair et al. (2010)) and eigensystem analysis are some of the most frequently used measures that fall into the second (for a thorough analysis see Section 1.1.1). All these measures commonly use some sort of rule of thumb to rule about the existence of multicollinearity. For each measure, at least 2 or even 3 different thresholds can be used; for instance, in the case of VIF 5, 10, and 20 are considered proper thresholds (see (Gujarati and Porter (2008), Wooldridge (2014) and Greene (2002), respectively). The question remains though: at which point extreme multicollinearity is actually extreme? All these methods usually fail to recognize patterns among variables due to weak or absent coefficients' penalization that results in variable over-elimination. So, how can someone properly address multicollinearity without risking increasing a models' bias that the omitted over-eliminated variables might cause? There is always a thin line between the worthiness of variable reduction, on one hand, and the robustness and validity of the results on the other. For a thorough discussion see Lindnee et al. (2020).

To resolve the issue, regularization techniques are used that are considered optimal for parsimonious model creation when an immense number of variables is involved. These techniques are based on beta coefficients penalization and aim to reform the coefficients as more unbiased as they can be by assigning weights ("of significance") that punish the insignificant or the less significant variables while simultaneously rewarding the statistically significant ones. Ridge (Tikhonov (1943),(1963)), Lasso (Tibshirani (1996)), and their aggregation, Elastic Net (Hastie et al. (2001), Zou and Hastie (2005)) are the most frequently used regularization approaches for addressing this issue. The disadvantage of these methods is that they can be computationally time-consuming.

[1]The results of this Chapter have been published as:
Ntotsis, K., Karagrigoriou, A. and Artemiou, A.: Interdependency Pattern Recognition in Econometrics: A Penalized Regularization Antidote, Econometrics, 9, 44, 2021.

In this work, a criterion is proposed based on the combination of penalized coefficients; more precisely we propose the generation of a criterion that combines penalized beta coefficients with a penalized coefficient of determination, both emerging from the naive Elastic Net and aims to enhance the generalizability of a learned model. The proposed criterion, namely Elastic Information Criterion (EIC), can be considered as a non-time or space consuming algorithmic procedure, which is more accurate than standard measures when it comes to pattern recognition among multicollinear variables. Another distinct characteristic of EIC is that it evaluates the existence and the magnitude of multicollinearity based on a unique data-driven threshold which is reckoned based on data peculiarities and not some approximate rule of thumb that typical measures rely on. The proposed criterion is expected to play the role of a supplementary tool in the hands of the researcher to be used in conjunction with their judgement, experience, and knowledge, together with any special characteristic associated with the problem/dataset at hand.

A motivating example

In this subsection, an example based on 3 random variables X_1, X_2, X_3 is used as a motivation for the proposed methodology. X_1 and X_3 are random samples of size $n = 100$ from the standard normal distribution, while X_2 is calculated as a function of X_1 through the expression

$$X_2 = u \times X_1 + \sigma \times \epsilon \tag{6.1}$$

where u is either 2 or 5, $\epsilon \sim \mathcal{N}(0, 1)$, and σ a constant that controls the variability of errors. For σ we use values in the set $\{0.2, 0.5, 1, 2, 5\}$. At the same time, u has been chosen to provide an additional, more general, interdependence structure between the variables involved. The example involves 10 datasets, each containing a unique combination of values for u and σ. This example seeks to see the efficiency rate of EIC and VIF, meaning how many times each measure manages to do proper variable selection, i.e., to select X_3 and either X_1 or X_2 variable. Note that in all cases X_3, due to its congenital randomness, never exhibits multicollinearity despite the measure chosen, and hence its interpretation is omitted, without indicating its ejection from the procedure. Table 6.1 provides the results of 1000 replications of the above experiment. In Table 6.1, it can be observed that the efficiency rate of VIF

measure / {u,σ}	EIC	VIF	Correlation Range
{2,0.2}	45%	0%	[0.98, 1]
{2,0.5}	40%	0%	[0.94, 0.98]
{2,1}	24%	1%	[0.78, 0.94]
{2,2}	16%	-	[0.46, 0.83]
{2,5}	7%	-	[-0.1, 0.59]
{5,0.2}	50%	0%	[0.99, 1]
{5,0.5}	67%	0%	[0.98, 1]
{5,1}	72%	0%	[0.96, 0.99]
{5,2}	70%	0.1%	[0.86, 0.96]
{5,5}	35%	-	[0.45, 0.83]

TABLE 6.1: EIC and VIF efficiency rates comparison for the motivating example for all u and σ combinations.

(based on a threshold value equal to 5) is excessively inadequate. More specifically, it does not make proper variable selection in at least 99 percent of cases. Additionally, there were cases of $\{u,\sigma\}$ ($\{2,2\}$,$\{2,5\}$,$\{5,5\}$) that multicollinearity was not detected by VIF. Given the prior knowledge that X_2 is indeed a figment of X_1, one can conclude that multicollinearity is lurking behind the generated randomness. Moreover, if the methodology to be proposed and presented in the sequel is applied in the motivating example, the results appear to be remarkable. Indeed, the efficiency rate of EIC is as high as 72% and, in any case, clearly prevails over VIF regarding variable over-elimination. Note that the corresponding rates for VIF were almost 0% or non-existent, meaning that in all replications both X_1 and X_2 appeared as multicollinear. The correlation range indicates the minimum and the maximum correlation between X_1 and X_2 of each dataset. More precisely, for each $\{u,\sigma\}$ combination, the experiment was replicated 100 times and the minimum and maximum correlation values between the variables, were registered. Among all experiments and all $\{u,\sigma\}$ combinations, the overall minimum and the overall maximum correlation values were used to provide the correlation range. The aim was to evaluate the performance of each measure under different degrees of correlation. Even though high correlations were detected in most cases (implying the possible existence of multicollinearity), VIF failed either to recognize it or detect it without being able to identify the predetermined pattern between X_1 and X_2. The example reveals a weakness of the VIF associated with its failure to identify patterns exhibited by the variables involved. The development of EIC came out of a necessity to fill this gap in the literature; i.e., to provide a measure capable not only to recognize multicollinearity patterns that lurk behind variables but also to work simultaneously, as a variable selection criterion.

6.1 Elastic Information Criterion

6.1.1 The penalized regularization antidote

In this Section, the Elastic Information Criterion (hereafter EIC) is proposed. EIC can be considered an extension of the Elastic Net procedure and result in a (computational) time and space non-consuming algorithmic procedure that has also proven to be more accurate than typically used measures regarding pattern recognition among multicollinear variables. The Elastic Net was selected as the optimal regularization due to its capability to examine the impact of different α_{en} and λ_{en} combinations on the model through a cross-validation procedure. EIC initiated out of necessity for accurate and effective multicollinearity capture without having variable over-elimination. Its aim is to detect patterns among the multicollinear variables and more precisely, which one enacts as a function of the other(s), and remove them, leaving the one(s) that originated from them intact. The EIC's results emanate from the Elastic Net cross-validation procedure, and its formula is given in the following form:

$$EIC_j = \alpha_{j,en} \times \frac{\sum\limits_{\substack{p=1 \\ p\neq j}}^{k} |\beta^j_{p,en}|^{1+\alpha_{j,en}}}{1 - R^2_{j,en}} \geq 0, \quad j = 1, 2, \ldots, k \tag{6.2}$$

and

$$X_j = \beta^j_{0,en} + \sum_{\substack{p=1 \\ p \neq j}}^{k} \beta^j_{p,en} X_p \tag{6.3}$$

where

- k is the total number of regressors (explanatory variables),
- $\alpha_{j,en}$ is the optimal alpha emerging from the Elastic Net procedure and corresponds to the modelling of the X_j variable,
- $\beta^j_{0,en}$ is the intercept term in Equation 6.3,
- $\beta^j_{p,en}$ is the penalized coefficient of the p^{th} regressor in Equation 6.3,
- $R^2_{j,en}$ is the R^2 of the j^{th} variable as predictor regressed against all other regressors.

EIC integrates two aspects of collinearity detection. The primary one, based on a tolerant method alteration, which aims to reduce the sensitivity of coefficients throughout the penalty function. The number of $\beta^j_{p,en}$ coefficients diversifies from zero to k since when $\alpha_{j,en}$= 1, then the variable's coefficient reduces to zero. The summation of this function aggregates all the resulting $\beta^j_{p,en}$ coefficients emerging through Elastic Net regression. On the other aspect, the goodness of fit in the linear model is used as a penalty for multicollinearity disclosure. Lastly, the tuning parameter $\alpha_{j,en}$ is utilized for penalization smoothing purposes. EIC tends to perform more precisely for $\alpha_{j,en}$ at or close to the end-point of the [0,1] range. Thus, in order to limit -in terms of time- the computational burden for $\alpha_{j,en}$ selection, the values examined range from 0 to 0.1 with step 0.01, the middle point of the $\alpha_{j,en}$ range (0.5), and from 0.9 to 1 with step 0.01. Note that otherwise the $\alpha_{j,en}$ specification, the same Cross-validation procedure as in the naive Elastic Net, is followed.

Algorithm 5 Pseudocode for EIC implementation in R

Input: A $n \times k$ matrix, namely A, containing the dataset with each X_j column representing a variable.

Output: A $1 \times k$ data frame containing the EIC value for each X_j variable indicating the level of multicollinearity.

Procedure: Compute $a_{j,en}$, $\beta^j_{p,en}$, and $R^2_{j,en}$ parameters of *Equation* 6.2 for each X_j variable

Step 1: Set the vector of the considered values alpha ($\alpha_{j,en}$), namely `alpha.sample <- c(seq(0, 0.1, by = 0.01), 0.5, seq(0.9, 1, by = 0.01))`.

Step 2: Perform `cva.glmnet` function, which is a part of `glmnetUtils` package, by setting the following arguments: `x = A[, -1], y = A[, 1])` and `alpha = alpha.sample`.

Step 3: The resulting arguments are as follows: `alpha` is the $a_{1,en}$, `lambda` is the $\lambda_{1,en}$, and $\beta^{j}_{p,en}$ are the penalized coefficients of the explanatory variables of the model considered.

Step 4: Compute the absolute value of each of the resulting $\beta^{j}_{p,en}$ coefficients raised to the power of $1 + \alpha_{1,en}$.

Step 5: Sum all the values resulted through **Step 4** in order to calculate the numerator of *Equation* 6.2.

Step 6: Compute the R^2_1 of the X_1 variable regressed against every other variable in the dataset which corresponds to the $R^2_{1,en}$ of *Equation* 6.2, based on the coefficients as resulted through **Step 3**.

Step 7: Replace the result of **Step 3-5** on *Equation* 6.2 and then calculate the EIC_1 value, which corresponds to the multicollinearity level of the X_1 variable.

Step 8: Repeat **Steps 1-6** for the remaining k variables.

6.1.2 Data-driven threshold

To verify the presence of multicollinear variables with EIC, the following threshold determined by the collection or analysis of data has been proposed.

$$threshold = \bar{\lambda}_{en} + 3 \times s.e.(\bar{\lambda}_{en}) \tag{6.4}$$

where $\bar{\lambda}_{en} = \frac{\sum_{j=1}^{k} \lambda_{j,en}}{k}$ and $s.e.$ stands for the standard error (of the sample mean $\bar{\lambda}_{en}$.) Adding three standard errors to the threshold, which is a typical quality control bound, reduces the possibility of wrongfully variable rulings.

Given a dataset of k variables and based on Equation 6.2 and Equation 6.4, one can conclude that a variable does not display multicollinearity for values of EIC lower than the threshold:

$$0 \leq \alpha_{j,en} \times \frac{\sum_{\substack{p=1 \\ p \neq j}}^{k} |\beta^{j}_{p,en}|^{1+\alpha_{j,en}}}{1 - R^2_{j,en}} \leq \bar{\lambda}_{en} + 3 \times s.e.(\bar{\lambda}_{en}) \tag{6.5}$$

Algorithm 6 Pseudocode for the *threshold* of EIC in R

Input: A $n \times k$ matrix, namely A, containing the dataset with each X_j column representing a variable.
Output: A single number which serves as *threshold* for ruling about the existence of multicollinearity.

Procedure: Compute Equation 6.4 for the input dataset

Step 1: The implementation of **Steps 1** and **2** of Algorithm 5 will result in the $\lambda_{1,en}$ which corresponds to the X_1 variable.

Step 2: The completion of Algorithm 5 will produce the values of $\lambda_{1,en}, \lambda_{2,en}, \ldots, \lambda_{k,en}$ parameters. Calculate their arithmetic mean.

Step 3: Find the standard error of the mean via the function `std.error` and triple the result.

Step 4: Sum the values resulted from **Step 2** and **3** to form the threshold value of Equation 6.4.

Remark: The proposed criterion resolves a defect in classical diagnostic measures, like VIF, by being capable of detecting interdependency patterns among variables. In that sense, it provides a powerful and supportive tool in econometric analysis, which is expected to complement effectively all other aspects (purpose of the study, researcher's judgement, etc.) of the decision-making process.

6.2 Numerical Applications

There are continuous and recurrent discussions in econometrics, regarding the way to effectively address the issue of multicollinearity. It is believed that, to some extent, this is due to the absence of simulated studies and the fact that in real cases, available data are simple and direct, which prevents an in-depth understanding of the issue, when in fact econometric research is considered particularly complex. In this research area, variables tend to be interdependent, while sample sizes are relatively limited. Therefore, due to the nature of the problem, it is difficult to have an interpretable application in real data. In order to investigate the validity of EIC, a real case scenario based on a dataset on the economic growth of a country's prosperity is presented bellow, followed by a simulated case study. In both experiments, a comparison concerning the proper variables' prediction rate, between EIC and various other measures has been implemented for evaluating the effectiveness of the proposed methodology.

6.2.1 Real case study

For validation purposes on real data, the following experiment was conducted. For evaluating a country's prosperity and having a better understanding of where its economy is headed, several economic growth indicators have developed throughout the decades. Some main closely monitored and widely applied indicators include the Balance of Trade to GDP (BoT), the Government Debt to GDP (GovDebt), the Gross Domestic Product Growth Rate (GDP_{GR}), the Inflation Rate (Inf_R), the Interest Rate (Int_R) and the Unemployment Rate ($Unem_R$). A dataset consisting of these 6 variables with annual observations covering the time period 2000 to 2020 for Greece was formulated for illustrating the performance ability of the proposal EIC criterion as compared with traditional diagnostic measures. Data originated from the OECD database (2021), the Trading Economics (2011), and the World Bank Open Data (2021). Based on the dataset, a direct interdependency pattern between GDP_{GR} and both GovDebt and BoT exists, since the latter two appear as percentages of the former. According to the relevant bibliography (see e.g. Dumitrescu et al. (2009), Fried and Howitt (1983), and Oner (2020)), correlations are observed between the variables involved in the dataset. The aim is (a) to observe whether the measures

mentioned in Section 1.1.1 can identify the aforementioned interdependency pattern among the variables, and (*b*) to observe how EIC corresponds to the same situation.

Individual Multicollinearity Diagnostic Measures

	EIC	VIF	TOL	CI	F-G w_j	Leamer	IND1
BoT	0	1	1	0	1	1	0
GovDebt	0	0	0	0	1	0	0
GDP_{GR}	1	0	0	0	1	0	0
Inf_R	0	1	1	1	1	0	0
Int_R	0	0	0	1	1	0	0
$Unem_R$	0	1	1	1	1	1	1

TABLE 6.2: Detection of existence (1) or not (0) of multicollinearity by diagnostic measures

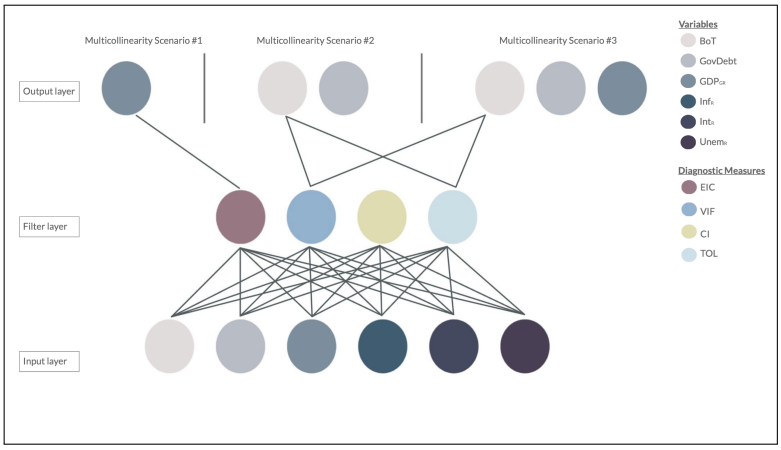


FIGURE 6.1: Neural Network-like infographic for diagnostic measures performance

In Table 6.2/Figure 6.1 for each variable, the existence (1) or not (0) of multicollinearity was detected by various diagnostic measures. As it can be seen, F-G w_j, Leamer, and IND1 measures perform poorly. Thus, the analysis will be focused solely on the comparison of EIC with VIF, CI, and TOL. In Figure 6.1 the performance of these measures can be seen in a neural network-like infographic. In the Input layer, we have the 6 variables of the dataset. In the Filter layer, the implementations of EIC, VIF, IC, and TOL can be seen in that order. In the Output layer, we have the three possible multicollinearity scenarios and expect the diagnostic measures to fall into one of them. Based on the above, the three scenarios are: (i) multicollinearity is detected only on GDP_{GR}, (ii) multicollinearity is detected on BoT and GovDebt; and (iii) multicollinearity is detected on all three, namely GDP_{GR}, BoT and GovDebt.

Based on Table 6.2/Figure 6.1 we observe that except EIC, the other measures, identify as multicollinear some of the variables BoT, Inf_R, Int_R, and $Unem_R$. However, the EIC restricts the multicollinearity issue solely to GDP_{GR} and identifies it as the "root" of multicollinearity in the dataset, as can been clearly seen in Figure 6.1 that falls into one of the three scenarios. It must be noted that the selection of this variable is of great importance due to its linkage to all others, and because this connection goes undetected by all other measures. On the other hand, the results clearly show that classic diagnostic measures, like VIF, fail to recognize the underlying pattern among the variables involved. On the other hand, the proposed EIC criterion

not only exposes the pattern but also identifies its root and recommends, correctly, its removal from the dataset.

This example clearly shows that EIC succeeds in identifying interdependency patterns when all other diagnostics measures fail. If such patterns are non-existent, all measures are expected to behave equally well. The superiority of the proposed criterion lies in the fact that it offers a powerful tool for pattern identification, which could be useful for researchers.

6.2.2 Simulation case study

This study is based on data generated from a standardized normal distribution with different scenarios, sample sizes, and number of variables. The number of variables ranged from 5 to 15, while the number of observations was 10, 50, and 100. Each scenario was replicated for validation purposes, providing similar results in all cases. Based on the similarity of the results, the decision to present the results for the same number of variables (10) and the same size of observations (100) throughout the study was made for comparability purposes.

The study focuses on three datasets with different degrees of correlation among variables, 20%, 45% and 75% for datasets "low", "medium" and "high" respectively. For each dataset, a sized 100×10 data frame was created and replicated 5000 times, with each X_j, $j = 1, 2, \ldots, k$, column representing a variable. For each dataset, several variables have been selected to be altered and involved in the analysis as linear operators of X_1 with the subsequent formula:

$$X_j = u \times X_1 + \sigma \times \epsilon, \quad X_j \neq X_1, \tag{6.6}$$

where u is a random number in $\{1, 2, 3, 4, 5\}$, $\epsilon \sim \mathcal{N}(0, 1)$ and σ is a constant that controls the variability of errors. For σ we use values in the set $\{0.2, 0.5, 1, 2, 5\}$. As in the case of the motivating example (Section 6), u has been chosen to provide an additional, more general, interdependence between the variables involved. Equation 6.6 was formulated out of necessity for implementing a more general interdependency pattern among the variables involved. Simultaneously, there was a need to explore the capabilities of the proposed methodology under a more challenging underlying mechanism (as opposed to the case of a fixed value for the u coefficient) for the building of the model in Equation 6.6. The selected linear transformations of X_1 are: X_2 for the low, X_2, X_3, X_4, X_6 and X_8 for the medium, and all X_j except X_8 and X_{10} for the high correlation-based category.

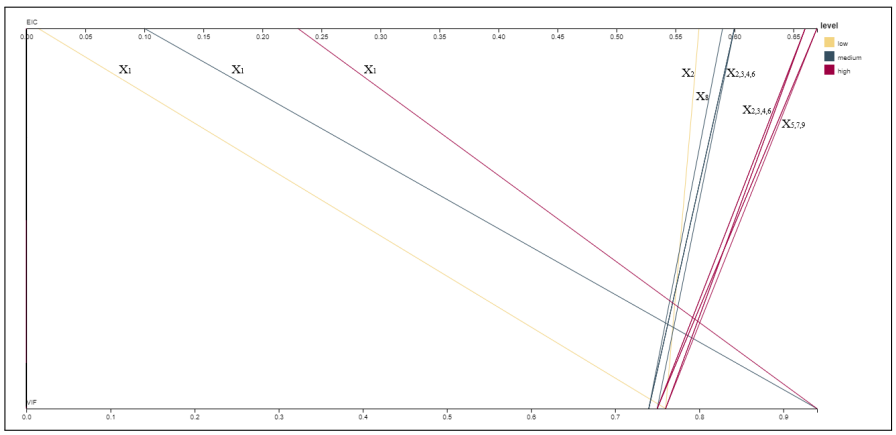


FIGURE 6.2: Parallel coordinates graph of EIC vs VIF

In Figure 6.2 we can observe the proportion of times each variable appears as multicollinear based on VIF (lower line) and EIC (upper line) in three (low/yellow - medium/green - high/red) correlation-based categories. Note that only the variables with non-zero proportions are displayed.

Note that the EIC was chosen to be compared only with the VIF, which is considered the most widely used diagnostic measure. A random sampling between the replications was held and the detection process with all measures was implemented, which did not display noteworthy results and verified the above-claimed decision.

The Parallel Coordinates Graph of Figure 6.2, which was carried out in RAW-Graphs (Mauri et al. (2017)), provides the percentage of times each X_j variable appears as multicollinear based on EIC (upper line) and based on VIF (lower line) with yellow lines corresponding to low, green to the medium and red to the high correlation-based dataset. High values (close to 1, i.e., 100%) indicate extreme multicollinearity, while low values (close to 0, i.e., 0%) indicate weak (or absence of) multicollinearity. As an example, consider the yellow line (low correlation dataset) associated with the variable X_1 (which has been taken to be non-multicollinear). The EIC correctly identifies the non-multicollinearity of X_1 since the upper line is crossed at a value less than 0.05 (the actual value is 0.01). Meanwhile, VIF fails to identify the same. Indeed, although the yellow line should have been vertical (crossing the lower line at about the same value as the upper line) the crossing is observed far to the right, at a value between 70% and 80% (the actual value is 0.76) indicating that VIF characterizes, incorrectly, X_1 as multicollinear.

Based on the above observations according to Figure 6.2, we can conclude that only EIC succeeds in correctly identifying the level of multicollinearity of all variables involved with X_1 appearing on the left corner (of the upper line of Figure 6.2) and all others on the right corner. We also observe that as correlation increases (from yellow to red), VIF is deceived and fails to recognize the unaltered variable (X_1) but instead, it signifies it, falsely, as the most multicollinear variable, which may result in variable over-elimination and improper model selection.

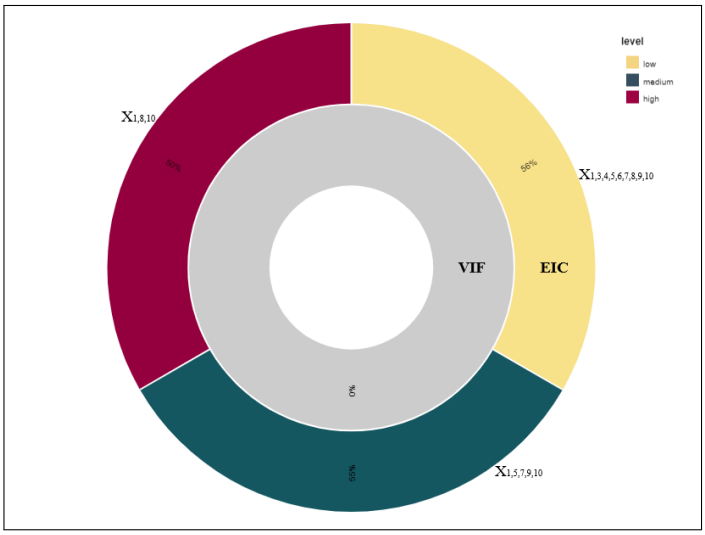


FIGURE 6.3: Proper model selection based on EIC and VIF for all three correlation-based categories.

In the sunburst diagram (Mauri et al. (2017)) of Figure 6.3, one can see the percentage rate at which each measure (VIF in the inner circle and EIC in the outer

circle) managed to properly do correct variable selection in each of the three categories. In Figure 6.3, VIF tends to do variable over-elimination and by expansion model misspecification. When the proper variables have been selected (all X_j except X_2 for low, X_1, X_5, X_7, X_9 and X_{10} for medium, X_1, X_8 and X_{10} for high), then all the other (improper) variables have been selected too. Thus, one can state that the accuracy rate of proper variable selection based on VIF is 0%. On the contrary, the equivalent rate based on EIC surpasses 50% in all cases.

6.3 Conclusions

Conclusively, the suggested Elastic Information Criterion procedure results in a robust and easily interpretable methodology for handling multicollinearity along with the appropriate data-driven threshold. The criterion constitutes a novel shrinkage and selection method since it is based on both the coefficient of determination and beta coefficients penalization, emerging in virtue of a biased (towards the endpoints of the mixing parameter α) Elastic Net, while the threshold has been established based on $\lambda_{j,en}$ tuning parameter of the same procedure. Thus, EIC is governed by the same or similar properties as those of Elastic Net. Additionally, it demonstrates a sufficiently sparse representative model with an adequate proper variable prediction rate, while firmly encouraging a grouping effect even when the significance of a variable is relatively limited.

The results of the real and simulated data analysis strongly suggest implementing EIC not only for econometric modelling and forecasting but also for classification purposes due to its high efficiency rate. EIC does not commonly fail with highly correlated data as opposed to typically used measures for multicollinearity detection, while its high prediction accuracy is due to the restricted values of the parameter α. Furthermore, EIC tends to perform better when the Elastic Net procedure is implemented at or near the $\alpha_{j,en}$ edges while it appears to have a robust variable selection accuracy rate over both real and simulated case studies. The pivotal characteristic of reduction or ejection of the insignificant coefficients that Elastic Net attains, manages to enhance its efficiency rate. In comparison to other multicollinearity detection measures, it is evident that EIC prevails in terms of proper variable selection accuracy. An additional finding of this work is that the implementation of EIC can be vital in the field of Econometrics, where interrelationships among variables frequently occur. Its capability to identify where (in which variable(s)) the troublesome multicollinearity lurks and penalize it accordingly minimizes a models' bias without resulting in variable under or over-elimination.

EIC, as a criterion for implementing the EN mechanism, is particularly effective in tackling multicollinearity that lurks behind variables (Hastie et al. (2001), Zou and Hastie (2005)) . Indeed, as displayed above in all levels and as compared with the most widely used measures, EIC (a) identifies the existence of patterns among variables, (b) is capable of recognizing and "selecting" the altered variables, leaving the unaltered ones intact and (c) achieves extreme values in the presence of perfect multicollinearity and also in the total absence of it. Based on these characteristics and properties we can say that the effectiveness of EIC can place it high in the list of measures that can be used to address the multicollinearity issue and in that sense it can be considered as a useful and effective tool in the hands of the researcher to be used in conjunction with their judgement, experience and knowledge together with any special features associated with the problem/dataset at hand.

In addition to the contributions of the proposed criterion to the multicollinearity literature, another advantage of EIC is that it operates as a variable/model selection criterion and consequently it can be exploited as a dimension reduction technique. It should be reminded, that these classical dimension reduction techniques, suffer from the fact that each generated component is a combination of different proportions of the original variables; thus it is often difficult to interpret the results (Zou et al. (2006)). On the other hand, the proposed EIC criterion manages to preserve the interpretability of the original variables because it relies simultaneously on shrinkage and sparse selection.

Chapter 7

A Comparative Study of Multivariate Analysis Techniques for Highly Correlated Variable Identification and Management[1]

The purpose of this work is to locate and analyse the interrelationships between GDP and various variables which are interdependent and often characterized by a high degree of multicollinearity. The OECD define GDP as "*the standard measure of the value added created through the production of goods and services in a country during a certain period. As such, it also measures the income earned from that production, or the total amount spent on final goods and services (less imports). While GDP is the single most important indicator to capture economic activity, it falls short of providing a suitable measure of people's material well-being for which alternative indicators may be more appropriate. This indicator is based on nominal GDP (also called GDP at current prices or GDP in value) and is available in different measures*" (OECD (2019)). Based on well-established and proven studies, it is known that GDP can be expressed by

$$GDP = C + I + G + (Ex - Lm) \tag{7.1}$$

where C represents the Private Consumption Expenditures, I the Private Domestic Investments, G the Government Consumption Expenditures, Ex the Total Exports and Im the Total Imports.

GDP is frequently used by central banks, public entities and private businesses as a standard measurement for the economic health of a country (Callen (2008)). For predictive purposes, researchers often rely on economic or financial indices and model identification procedures. den Reijer (2005) and Schumacher (2007) both studied the forecasting of Dutch and German respectively, GDP through variable modelling. Later, Akhter et al. (2012) used PCA in order to obtain a model for the GDP of Bangladesh. Bai et al. (2015) have shown the accuracy of variable analysis in the evaluation of the economy of a country, including variables such as Unemployment Rate, Investments, Population and General Government Total Expenditures, which are part of the current model analysis. Because of its unstable economy, Greece is the focus of many economic analyses from organizations such as the OECD, Eurostat, International Monetary Fund, and there is sufficient material and data on their websites one can refer to.

[1]The results of this Chapter have been published as:
Ntotsis, K., Kalligeris, E.N. and Karagrigoriou, A.: A Comparative Study of Multivariate Analysis Techniques for Highly Correlated Variable Identification and Management, In- ternational Journal of Mathematical, Engineering and Management Sciences, 5(1), 45-55, 2020.

The explanatory variables (Table 7.1) that were chosen are highly correlated and result in severe multicollinearity in the primary model, which appears to be a frequent problem in financial and economic big data analytics (Wang and Alexander (2019)). For the reduction or even elimination of the multicollinearity, which is a common issue in data analysis in finance and economics (Kondo et al., (2018)), a number of dimension reduction techniques were used in order to identify an optimal model with a set of new uncorrelated variables/variables. In this work, for comparative purposes and for measuring the quality of each model, three information criteria were used, namely AIC (Akaike (1974), BIC (Schwarz (1978)) and MDIC (Mantalos et al. (2010)).

Exports of Goods and Services (X_1)	Investments (X_5)
General Government Total Expenditures (X_2)	Population (X_6)
Household Consumption Expenditures (X_3)	Total Labor Force (X_7)
Imports of Goods and Services (X_4)	Unemployment Rate (X_8)

TABLE 7.1: Explanatory variables

In this work we rely on multivariate analysis and in particular, on DRT for the modelling of the GDP by identifying an appropriate set of variables from a long list of possible explanatory interdependent variables which likely interact with and affect the GDP. The choice of GDP is obvious since it is a quantity of great interest for micro as well as macroeconomics. The case of Greece is chosen due to extreme economic events of recent years that greatly affected all aspects of economic activity.

7.1 Preference Data

Gross Domestic Product is interrelated, according to relevant theory, with a variety of explanatory variables which possibly affect GDP.

This work is based on Greece's economy with annual data collected through Knoema, OECD and Eurostat for the eight (8) explanatory variables X1–X8 presented in Table 7.1 for the period 1980- 2018 (39 annual observations). Three (3) missing values have been replaced by the average values of the preceding and the following year.

7.2 Dimension Reduction Techniques

Based on the overall results of the implementation of PCA and the fact that it is preferable to avoid the loss of important information, we conclude that the first two components (Z_1 and Z_2) should be kept (see Table 7.2) regardless of the eigenvalues because they retain a considerable amount of the total information/variability (more than 95% of the original variability of the data). The described variability played a key role in the aforementioned decision, since the intention was to keep that many components, so that a considerable proportion of the original variability will be described by the components chosen.

1st Component (Z_1)	**2nd Component (Z_2)**
General Government Total Expenditures (0.97)	Investments (0.62)
Household Consumption Expenditures (0.99)	Unemployment Rate (0.74)
Imports of Goods and Services (0.97)	
Total Labor Force (0.97)	

TABLE 7.2: The two primary PCs

Remark: To determine which variables were significant in each component, the following empirical rule was followed. For the two chosen components, the variables for which the absolute value of the associated coefficient is at least equal to 0.95 are kept as significant. A value of around 0.95, although there is no specific rule, is considered to be satisfactory in retaining a sufficient amount of information.

For the problem at hand, the first component, denoted by Z_1, holds more than 80% of the total variation of the dataset, while the second one, denoted by Z_2, holds roughly 15% of it. The rest of the components contain the remaining percentage of variation. By construction, the first component is considered to be the most important, on which the analysis is primarily based. Having said that, we observed in the above analysis, 6 of the total of 8 variables emerge as important according to the associated coefficients given in parenthesis (see Table 7.2).

Remark: For modelling purposes both PCA significant variables/components (Z_1 and Z_2) are used in their full form that contains, not only the significant variables (with coefficients at least equal to 0.95) which are presented in Table 2, but all m=8 original $X_i's$.

As it can be seen from Table 7.2, General Government Total Expenditures, Household Consumption Expenditures, Imports of Goods and Services and Total Labor Force emerge as important in the first component while Investments and Unemployment Rate in the second one.

Hence, using this technique we proceed with the Multivariate Analysis of the Gross Domestic Product with Z_1 and Z_2 as the uncorrelated variables affecting GDP.

The implementation of the $Merit_{S_N}$ procedure (Section 3.1.8) results in the withdrawal of 5 out of the total 8 original variables. The remaining variables, namely General Government Total Expenditures, Household Consumption Expenditures and Imports of Goods and Services are considered as the important ones in the modelling of GDP. It must be noted that the same variables together with the Total Labor Force compose the first and most important component (Z_1), of PCA.

7.2.1 Techniques Review

The aforementioned dimension reduction techniques were implemented for the identification of interrelationships between a number of potentially significant variables and GDP. While in some cases similarities between the techniques were revealed, all three highlight different variables as important, as it can be seen in Table 7.3.

Beale et al.	**PCA**	**CFS**
Total Labor Force	General Government Total Expenditures	General Government Total Expenditures
Unemployment Rate	Household Consumption Expenditures	Household Consumption Expenditures
	Imports of Goods and Services	Imports of Goods and Services
	Investments	
	Total Labor Force	
	Unemployment Rate	

TABLE 7.3: Variable selection based on examined criteria

7.3 Model Selection Criteria

Model identification procedures play a pivotal role in statistics by identifying the best model among an available class of models. Those techniques are contemplated as assessors of a quantity. For example, for a given data set the probability of the proposed model can be used as an assessor, which is essential for the pursuit of identifying the optimal fundamental structure of the phenomenon under investigation.

Model identification procedures have been heuristically recommended for time-varying processes. Kullback and Leibler (1951) developed such a measure that minimizes the loss of information. A direct connection between the Kullback-Leibler (KL) measure and the Maximum Likelihood Estimation (MLE) method, gave rise to AIC and BIC. In this work, we rely on AIC, BIC and MDIC to obtain the optimal model.

7.3.1 Model Selection based on AIC

In the previous section three-dimension reduction/variable selection techniques were used in order to find the optimal explanatory variables for the modelling of GDP, namely, Beale et al., PCA and Hall's CFS Selection technique. Using GDP as the dependent variable and the selected variables of each technique as the independent ones, the following three models were constructed corresponding to the aforementioned techniques respectively:

$$Y_i = \alpha_{11} + \beta_{11} X_{i7} + \beta_{12} X_{i8} + \varepsilon_{i1} \quad i=1,...,39$$
$$Y_i = \alpha_{21} + \beta_{21} Z_{i1} + \beta_{22} Z_{i2} + \varepsilon_{i2} \quad i=1,...,39$$
$$Y_i = \alpha_{31} + \beta_{31} X_{i2} + \beta_{32} X_{i3} + \beta_{33} X_{i4} + \varepsilon_{i3} \quad i=1,...,39.$$

The results based on MDIC in conjunction with those based on AIC and BIC are provided in Table 7.4

	AIC	BIC	MDIC
Beale et al.	2009	2015	5
PCA	1925	1941	37
CFS	1901	1909	7

TABLE 7.4: Model Selection Summary

From the results in Table 7.4, it appears that the optimal model based on AIC is the one formulated by Hall's CFS technique and contains the General Government Total Expenditures, the Household Consumption Expenditures and the Imports of Goods and Services as the independent variables.

Additionally, in Table 7.4 the AIC values range from 1900-2000 for all three methods, with CFS providing the best model. Based on BIC, we conclude the same outcome as AIC. However, as can be seen from the figure, MDIC provides by far the most optimal models with values ranging from 5 to 37. Based on this measure, Beale et al. and CFS provide the optimal models.

7.4 Conclusion and Future Research

In conclusion, in this work, we attempted via dimension reduction techniques, to identify interrelationships between the GDP of Greece and a number of variables which are highly correlated. Beale et al. (1967), PCA and Hall's CFS techniques were implemented and suggested different models with different variables (see Table 7.3).

More specifically, Beale et al. proposed a model with the Total Labor Force (X_7) and the Unemployment Rate (X_8) as independent variables. This technique clearly focuses solely on the workforce point of view in order to achieve the optimal model. PCA, on the other hand, instead of using the original variables, created new uncorrelated ones. In fact, PCA promotes a model with two uncorrelated variables (Z_1 and Z_2). Through them, 6 out of a total of 8 variables emerge as important, namely X_2, X_3, X_4, X_5, X_7 and X_8 (see Table 7.2). It should be noted that the variables selected as significant have also been chosen either by Beale's or Hall's models. The third technique, CFS, proposed a model with the General Government Total Expenditures (X_2), the Household Consumption Expenditures (X_3) and the Imports of Goods and Services (X_4) as significant variables affecting GDP.

Based on theoretical background (see Equation 7.1), it appears that the CFS model covers most part of GDP's formula and seems to be able to identify and select the "right" subset of variables from the original ones. Indeed, although CFS does not select the Investments and the Exports of Goods and Services which both are part of the variables involved in Equation 7.1, it is able to identify, the Imports of Goods and Services (which is part of the Imports), the Government Expenditures and the Household Consumption Expenditures. Note though, that the CFS model also chooses to ignore demographic variables, which affect indirectly and not directly the modelling of GDP through their interrelationships with all variables involved in Equation 7.1.

The theoretical interpretation of the results is confirmed by two out of the three model selection criteria that were used, and their results are provided in Table 7.4. Both AIC and BIC select Hall's CFS model, while MDIC selects Beale et al. model.

From the analysis, we see that the PCA model is not optimal in all cases examined. When it comes to CFS and Beale et al., we observe that, both AIC and BIC, choose clearly the former, leaving way behind the latter. On the other hand, although MDIC is in favour of Beale et al. (1967), the difference observed as compared to CFS, could not be considered significant.

The main obstacle that we had to overcome in this work was the problem of multicollinearity, which is very common, especially when it comes to modelling that involves big data on various financial characteristics and/or economic indicators. The case of the GDP of Greece was an ideal example to explore the capabilities of various multivariate analysis techniques in handling the multicollinearity problem and identifying a set of influential variables.

Taking that under consideration, it is possible, in a future work, to attempt to explore how different model selection criteria react or are able to make the right variable/model selection, when multicollinearity is of different magnitude. Through this process one could be able to identify the criterion which is better adjusted and finally succeeds in choosing the optimal model when the variables involved are highly correlated.

Chapter 8

Multistep Dimension Reduction for Credit Scoring Modelling[1]

A credit scoring model (*CSM*) is an effective and important mechanism that helps in maximizing the risk-adjusted return of a financial institution, an enterprise or even an individual. A credit scoring methodology is a type of statistical analysis for accessing one's creditworthiness, and as such it has to be as accurate as possible in terms of prediction. The interest in predictive as well as descriptive performance for scorecard construction is endless.

Credit rating modelling has been of great interest to Finance and Banking since as early as the 50s and the 60s. Various techniques have been used over the years based on logistic regression, discriminant analysis, support vector machines, neural networks etc. (Eisenbeis (1978), Hardy and Adrian (1985), Bellotti and Crook (2009), Chen et al. (2011), Yu et al. (2010), Boritz and Kennedy (1995), Kumar (2005), Paleologo et al. (2010), Mavri et al. (2008), Mavri and Ioannou (2004)).

The objective of this work is the proposal for descriptive (classification) as well as predictive purposes, of an innovative approach to flexible and accurate credit scoring modelling which is of significant importance in Finance and Banking due to its direct connection to one's creditworthiness.

For the development of a flexible and reliable forecasting modelling approach, we deal with binary regression type models with financial as well as credit behaviour data (Section 8.1.1) and focusing on the classification of businesses according to the risk of default (Basel Committee on Banking Supervision (2004)) into two classes for "good" (i.e., with "no delinquency") and "bad" (i.e., with "severe delinquency") credit behaviour (Section 8.1.2). Finally, the problem of dimension reduction in CSM is addressed by combining regularization methods and model identification techniques. The combination of financial and credit behaviour data (e.g., credit behaviour characteristics such as credit limit, current balance, frequency and amount of loan instalments, etc.) is quite original as most countries and institutions use only financial data for credit scoring modelling (Boguslauskas et al. (2011)).

For the modelling, we propose a 3(4)-step algorithmic procedure for dimension reduction with an initial preliminary data pre-processing step (*step0*). The latter is used for creating dummy variables using Weight-of-Evidence (WoE), a tool for measuring the degree of strength for separating bad and good enterprises. The main part of the algorithm is based on dimension reduction techniques taking into consideration a stepwise regression based on AIC (stepAIC) and a PCA. The proposed

[1]The results of this Chapter have been published as:
Giannouli, P., Karagrigoriou, A., Kountzakis, C.E. and Ntotsis, K.: Multilevel Dimension Reduction for Credit Scoring Modelling and Prediction: Empirical Evidence for Greece, Communications in Statistics: Case Studies, Data Analysis and Applications, 7(4), 545-560, 2021.

procedure allows for an optional 4th step based on ENR (Zou et al. (2005)) for further dimension reduction if the researcher feels that it is of use.

Through this study, we expect to succeed in (a) the categorization/classification of businesses (enterprises) into good and bad, regarding credit scoring, through statistically significant explanatory variables and (b) the selection of the optimal forecasting model for predicting the credit scoring of Enterprises.

8.1 Data Description and Pre-Processing

The proposed procedure is applied to the Greek system separately for Small and Large Enterprises/Businesses (according to their revenue - see Section 8.1.1) with data collected through a Credit Bureau operating as an inter-banking information system, in Greece. The analysis is based on a random sample of Greek enterprises collected through a Credit Bureau operating in Greece, for the period 1/1/2012 – 31/12/2014. The dataset consists of 88 variables with the following characteristics:

Enterprises	Number of Financial Variables	Number of Credit Behaviour Variables	Total Number of Observations
Small	27	12	73.661
Large	37	12	131.752

TABLE 8.1: Data characteristics for Small and Large Enterprises

8.1.1 Data Description

The data for the analysis in this work constitute a representative random sample of 4579 Greek businesses chosen from the database of Tiresias S.A., a Credit Bureau operating as an inter-banking information system in Greece, and have been recently analysed in Giannouli and Kountzakis (2019). Tiresias database contains financial data including credit-related data for individuals and all Greek businesses. The random sample used in this analysis consists of 1889 Small Businesses (with revenue at most 700,000 euros) and 2690 Large Businesses (with revenue at least 700,000 euros). It should be noted that enterprises (i) with insufficient history (of less than six months), (ii) that have chosen not to display their data in the system, and (iii) with negative behaviour (bad credit) during the observation period, have been excluded from the analysis.

In this work, as it is typical in such analyses (Siddiqi (2006)), a period of twelve months (from 01/01/2014 to 31/12/2014) is used as a performance period and a 2-year period as an observation period (from 01/01/2012 to 31/12/2013).

The purpose of the analysis is the modelling of a response variable representing business credit behaviour characterized as either “good” or “bad”. (For more details on the distinction between “bad” and “good” businesses, please see Section 8.1.2). The variables used for the analysis are divided into two main categories that correspond to

- **Financial Data** found in the enterprise’s balance sheet including the associated Financial Ratios (Barnes (1987), Boguslauskas et al. (2011)), and
- **Credit Behaviour Data** found in the Credit Consolidation System (CCS), the Default Financial Obligation System (DFO) and the Mortgages and Pre-notations to Mortgages System (MPS) like Delinquency Index, Credit Limit, Current

Balance, Current Balance Delinquent, Frequency of instalments (for loans), Amounts of instalments, Deletion flags etc., and variables emerged from them. For a full description of the variables used, see Giannouli and Kountzakis (2019).

In statistical terms, the scope of this work is the modelling of a binary classification problem for credit scoring. For the analysis, we will be using multivariate analysis techniques including logistic regression and model selection criteria for the identification of the most significant financial and credit behaviour variables for predictive purposes. The analysis is performed separately for Small and Large Businesses with the description of the selected variables given in Appendix B.

8.1.2 Credit Performance Characterization

The models proposed in this work intend to classify businesses' ratings according to the risk of default on their obligations. The proposed models include variables associated with the enterprise's both past and present financial behaviour. More specifically, the primary purpose of the proposed model is to discriminate businesses with a "bad" credit behaviour from those with a "good" credit behaviour based on the available data for the performance period.

The "bad" and "good" credit behaviour of a business is defined as follows (Basel Committee on Banking Supervision (2004), Siddiqi (2006)):

1. Businesses with a "Good" (positive) credit behaviour are businesses with **no delinquency**, businesses with either maximum delinquency from 0 to 29 days past due, during the last 12 months or with credit limit utilization over 102% for 0 to 29 days, concerning Small and Medium-sized Enterprises (SME) Overdrafts.

2. Businesses with a "Bad" (negative) credit behaviour are businesses showing **severe delinquency**, with either

   a SME Contracts (excluding Overdrafts) with maximum delinquency during the last 12 months, greater or equal to 90 days past due or

   b SME Overdrafts with maximum delinquency during the last 12 months, greater or equal to 90 days past due or credit limit utilization over 102% for time period greater or equal to 90 days with over limit amount greater than 100.

3. For the case of Guarantor, the characterization "Bad" refers to case (2) above with 150 instead of 90 days or more.

It should also be mentioned that an enterprise is characterized as having a "Bad" credit behaviour if during the performance period, a new DFO (loan denunciation) has occurred.

Finally, it is noted that the (credit) utilization of an enterprise is the ratio of the outstanding balance of the enterprise to its credit limit.

8.1.3 Step 0: Data Pre-processing

In order to prepare the data for the main part of the analysis in DR, we proceed in this section, into standard pre-processing operations by grouping the variables and/or

forming dummy variables with the use of WoE, also known as attribute strength, which represents the odds ratio and is defined by Siddiqi (2016):

$$WoE = ln\left(\frac{Distr.Good}{Distr.Bad}\right) * 100 = ln\left(\frac{Good/Total\ Good}{Bad/Total\ Bad}\right) * 100$$

where $Good/Total\ Good$ is the proportion of good businesses having a specific attribute and $Bad/Total\ Bad$ is the proportion of bad businesses having the same attribute. Such an approach is known to offer a number of advantages, including:

- an easy approach for outlying any infrequent observations and/or classes;
- easy to interpret relationships and explain their nature, therefore providing a better understanding of the phenomenon under investigation;
- non-linear dependencies can be analysed with the use of linear modelling techniques.

Such a procedure ensures that WoE will be sufficiently different among groups and hence, it will be expected to maximize the differentiation between bad and good enterprises, which is the primary objective of the analysis. Indeed, through such an analysis we focus on the identification of those characteristics that result in the best separation. Note that for this purpose, we rely on the WoE difference groups that play a key role in establishing and securing differentiation. As expected, larger differences resulted in higher predictive ability. It is also noted that WoE must also be in a logical order (i.e., in a ascending order from the worst to the best group categorization) for it to make operational sense.

8.2 Proposed Multistep algorithm

In data analysis, the first and most crucial problem that a researcher should overcome is the correct data interpretation. Indeed, whenever we deal with big datasets like the ones in this work, we are entering into the field of BDA, where the existence of collinearity is, among others, one of the most serious problems encountered associated with unreliable results.

During a preliminary analysis, various models, techniques as well as combinations of techniques have been considered for both datasets for Small and Large Enterprises, with the optimal combination resulting in a 3(4)-step algorithmic procedure consisting of the following:

Step 1 Data Standardization

Step 2 stepAIC

Step 3 PCA

Step 4 Elastic Net Regularization (*optional step*).

The purpose of the above algorithm is the dimension reduction which is achieved in two levels (in Steps 2 and 3): firstly, by the use of the stepAIC procedure applied to the standardized variables of step 1, and later by performing PCA in the variables selected by stepAIC. Based on the data and the final results of this study, an additional step (Step 4) is recommended to be included as optional, in the above algorithmic procedure. The use of this optional step is recommended if the data

justify its use. More specifically, after the 3-step algorithm is completed, a logistic regression analysis is performed using the variables selected in the later step of the procedure. The optional (4th) step can be considered as a 3rd level dimension reduction technique that removes, via ENR (Zou et al. (2005)), those PCs that do not contribute significantly to the proposed logistic regression model. The optimal models for both datasets (for Small and Large Enterprises) were selected based on two frequently used criteria, namely AIC and Adjusted R^2.

The proposed algorithmic procedure addresses and succeeds to resolve

a. the problem of multicollinearity and any other consequence of dealing with BD and

b. the limitation of the explanatory variables (*variables*) and, on one hand, making it possible to identify a flexible and easy-to-use model for predictive purposes and, on the other, a clear and precise interpretation of the results.

Step 1: Data Standardization

In standard data analysis, data standardization is often recommended before PCA. Indeed, if PCA is performed directly on the original explanatory variables, the newly emerged PCA variables fail to be (fully) independent, although this is the main goal of the implementation of PCA. This phenomenon may be attributed to heavy multicollinearity between explanatory variables with different measurement scales. In our analysis, we observed a high degree of multicollinearity as indicated by the Variance Inflation Factor (VIF; results not shown). In order to limit or eliminate it, data standardization was done, which affected considerably the correlations involved. After the first step of the procedure, multicollinearity in both datasets was observed to be significantly reduced although it still existed.

Step 2: Stepwise AIC

After the data standardization, the stepAIC procedure was applied as the first dimension reduction/feature selection criterion. This technique is one of the most common techniques used which attempts to identify the optimal subset of variables by minimizing the AIC value among the competing candidate models which given by AIC formula in Chapter 3. StepAIC has the ability to keep intact the larger possible part of the model's performance by simplifying it which results in the quantification of the amount of information loss. Note that, at each stage of the process, the technique checks whether variables that were removed in a previous phase become significant and are required to return to the model. For more, see Cavanaugh (2004), Shang and Cavanaugh (2008), Yamashita et al. (2007), and Zhang (2016)

Table 8.2 contains the results of Step 2 while presented the selected variables to be used in Step 3 of the analysis.

Small Enterprises	**Large Enterprises**
AIC = 1165	AIC = 1215
R^2 = 50%	R^2 = 50%
Adjusted R^2 = 50%	Adjusted R^2 = 49%
9 Financial Variables	**10 Financial Variables**
Debt Equity Ratio	Cash Ratio
Return on Equity	Current Assets to Total Liabilities
Working Capital Leverage	Net Profit Margin
Total Assets Turnover Ratio	Current Liabilities Turnover Ratio
Return on Assets	Fixed Assets to Equity
Total Liabilities	Working Capital Turnover Ratio
Short-term Liabilities	Total Liabilities
Result Curried Forward	Long-term Liabilities
Profit Before Taxes Depreciation and Amortization Expense	Total Fixed Assets
	Short-term Liabilities
6 Behaviour Variables	**8 Behaviour Variables**
Worst Payment Status in Last 3 Months	Maximum Utilization Not Revolving
Maximum Number of Months Consecutive with Over 100% Utilization in Last 6 Months	Worst Payment Status Last Month vs Last 24 Months
Number of Occurrences with Delinquency 1+ in Last 12 Months	Worst Payment Status in Last 3 Months
Maximum Number of Months Consecutive with Over 100% Utilization in Last 24 Months	Maximum Number of Months Consecutive with Over 100% Utilization in Last 6 Months
Current Balance/Delinquency to Current Balance	Number of Occurrences with Delinquency 1+ in Last 12 Months
Worst Payment Status Last Month vs Last 24 Months	Maximum Number of Months Consecutive with over 100% Utilization in Last 24 Months
	Total Current Balance
	Current Balance/Delinquency to Current Balance

For variables interpretation see Appendix B

TABLE 8.2: Small and Large Enterprises model selection summary - stepAIC

After the implementation of the stepAIC procedure, R^2 and Adjusted R^2 for both Small and Large Enterprises remain unchanged as in the original full model. Nonetheless, a noteworthy decrease in the AIC value can be observed in both cases. The AIC of the full model drops from 1200 to 1165 and from 1235 to 1215 after the implementation of the first -stepAIC- the DRT, in Small and Large Enterprises, respectively. Additionally, even if the AIC resulted in the same values for the full and the stepAIC models, the second one would be preferred due to its simplicity. One substantial dexterity of StepAIC is that the resulted models contain approximately only 38% (Small) and 36% (Large) of the variables used in the full model. As a result, the proposed models are more flexible and thus, preferable for predictive purposes, than the original ones based on AIC.

Step 3: Principal Component Analysis

The 2nd level dimension reduction procedure is applied to the 15 and 18 explanatory variables (see Table 8.2) selected by stepAIC for the Small and Large Enterprises, respectively. For this purpose, the classical PCA technique based on the correlation matrix is used as the second dimension reduction technique

In both datasets under consideration, we choose to retain the components that interpret approximately 90% of the overall variability of the original (standardized) variables. It is noted that various scenarios were studied with 80% and 75% variability as well as Kaiser's rule. The model with components (and by extension the

PCA variables) interpreting 90% of the total variability was the one for which the AIC and Adjusted R^2 values coincide with the corresponding values of the model obtained by stepAIC at the end of step 2 of the process. Although there is no specific rule to identify the statistically significant variables for each V_i, a proportion is considered to be satisfactory when it is able to retain a sufficient amount of the original information (Ntotsis et al. (2019)). Please note that this process simplifies the model (by reducing the number of PCA variables) without sacrificing the validity and effectiveness of the proposed model.

Based on the above Remark and in order to explain at least 90% of the total variability, for the Small Enterprises we retain only the first 9 out of 15, V_i variables while for the Large Enterprises we retain the first 11 out of 18, V_i variables. For forecasting purposes, the logistic regression will be applied to both datasets, using the model with the 90% variability. The results including the coefficient estimates for both cases under investigation are presented in Tables 3 and 4.

Small Enterprises Regression				
	Estimate	**Std. Error**	**test value**	**p-value**
(Constant)	0.3155109	0.0075925	41.556	< 2e-16
V_1	-0.0662298	0.0016387	-40.416	< 2e-16
V_2	-0.0406558	0.0031273	-13.000	< 2e-16
V_3	0.0307734	0.0045453	6.770	1.71e-11
V_4	0.0187028	0.0057085	3.276	0.00107
V_5	0.0049364	0.0066315	0.744	0.45673
V_6	0.0001217	0.0094362	0.013	0.98971
V_7	-0.0051696	0.0118533	-0.436	0.66279
V_8	0.0333873	0.0160883	2.075	0.03810
V_9	-0.0375664	0.0199572	-1.882	0.05994
MSE:	0.33	**AIC:**		1184.08
Multiple R-squared:	0.4984	**Adjusted R-squared:**		0.496

TABLE 8.3: step 3 - Small Enterprises regression and AIC results

Large Enterprises Regression				
	Estimate	**Std. Error**	**test value**	**p-value**
(Constant)	0.238662	0.005891	40.515	< 2e-16
V_1	-0.045877	0.001012	-45.323	< 2e-16
V_2	0.041277	0.002426	17.018	< 2e-16
V_3	0.032956	0.003291	10.015	< 2e-16
V_4	-0.027626	0.004825	-5.725	1.15e-08
V_5	0.033111	0.005120	-6.467	1.18e-10
V_6	-0.018376	0.006501	-2.827	0.00474
V_7	-0.018376	0.007557	-1.432	0.15239
V_8	-0.023566	0.009055	-2.603	0.00930
V_9	-0.031837	0.010407	-3.059	0.00224
V_{10}	0.026516	0.011492	2.307	0.02111
V_{11}	-0.049651	0.013419	-3.700	0.00022
MSE:	0.30	**AIC:**		1268.686
Multiple R-squared:	0.4886	**Adjusted R-squared:**		0.4865

TABLE 8.4: step 3 - Large Enterprises regression and AIC results

One can see that AIC and Adjusted R^2 values are very close to the corresponding values of the model selected with stepAIC prior to PCA implementation for both categories of enterprises (see Table 8.2). In other words, both models selected at the end of step 3 of the algorithmic procedure are much simpler than the ones selected in step 2 and at the same time retain a considerable amount of information. Hence, the second dimension reduction approach in step 3 chooses a number of proper variables for each class of enterprises (with 9 and 11, respectively, for Small and Large Enterprises). Meanwhile, no significant alteration in Adjusted R^2 and AIC results occurred compared to the full PCA models (with 15 and 18 variables, respectively).

Step 4: Elastic Net regularization – optional DRT

After the dimension reduction is completed, the final model is obtained by using a logistic regression analysis separately for Small and Large Enterprises using respectively, the 9 and 11 variables selected through the proposed algorithmic procedure.

Taking into account the regression results in Table 8.3 and Table 8.4, we can move on to an optional third level of dimensionality reduction, as long as the results allow it. Specifically, the results extracted through logistic regression revealed a number of statistically non-significant PCA variables (e.g., at significance level $\alpha = 5\%$). The reduction in the number of variables combined with the fact that the removed variables are statistically insignificant often results in models with a better AIC due to a lower penalty term. In order to ratify the above observation, an ENR was implemented.

The implementation of ENR reveals that in the case of Small Enterprises, the PCA variables V_6 and V_7 are statistically non-significant (a result also confirmed by the Student's t-test). The final proposed model, which can be used for predictive purposes, given in Table 8.5, has a better AIC than that of step 2 of the procedure and includes 7 PCA variables (with 15 initial - standardized variables each).

Small Enterprises Regression				
	Estimate	**Std. Error**	**test value**	**p-value**
(Constant)	0.315511	0.007588	41.581	$<$ 2e-16
V_1	-0.066228	0.0016387	-40.449	$<$ 2e-16
V_2	-0.040645	0.003125	-13.005	$<$ 2e-16
V_3	0.030745	0.004542	6.769	1.73e-11
V_4	0.0186688	0.005705	3.272	0.00109
V_5	0.033111	0.005120	-6.467	1.18e-10
V_8	0.033330	0.016078	2.073	0.03831
V_9	-0.037544	0.019945	-1.882	0.05994
MSE:	0.32	**AIC:**		1180.042
Multiple R-squared:	0.4983	**Adjusted R-squared:**		0.4967

TABLE 8.5: step 4 - Small Enterprises regression and AIC results

For Large Enterprises in Table 8.4 only V_7 is statistically non-significant based on t-test and also confirmed by ENR. It can also be seen from Table 8.6 that the resulted model retains its credibility since it has the same AIC value as the model selected in step 2, although it has one less PCA variable.

***Remark 2**: The procedure of the optional step 4 technique is applicable, provided that there is at least one statistically significant variable in the final logistic regression model. In this particular case study, the contribution for both cases could be considered relatively*

Large Enterprises Regression				
	Estimate	**Std. Error**	**test value**	**p-value**
(Constant)	0.238662	0.005892	40.515	$<$ 2e-16
V1	-0.045878	0.001012	-45.314	$<$ 2e-16
V2	0.041302	0.002426	17.025	$<$ 2e-16
V3	0.032987	0.003291	10.023	$<$ 2e-16
V4	-0.027586	0.004826	-5.716	1.21e-08
V5	0.033181	0.005120	-6.480	1.09e-10
V6	-0.018380	0.006502	-2.827	0.00474
V8	-0.023551	0.009057	-2.600	0.00936
V9	-0.031751	0.010409	-3.050	0.00231
V10	0.026567	0.011494	2.311	0.02089
V11	-0.049665	0.013422	-3.700	0.00022
MSE:	0.30	**AIC:**		1268.743
Multiple R-squared:	0.4882	**Adjusted R-squared:**		0.4863

TABLE 8.6: step 4 - Large Enterprises regression and AIC results

limited since the comprehensive dimensionality is reduced by three (dimensions) which also contribute to the improvement of the overall performance of the model.

It is worth mentioning that the proposed models are quite useful and more effective than models based on financial data only. Indeed, both the adjusted R^2 and the model selection criteria verify the superiority of the combination of financial and credit behavioural data by exhibiting a considerable improvement as compared with the models based exclusively on financial data.

8.3 Conclusions

The objective of this work is the proposal for descriptive (classification) as well as predictive purposes, of an innovative approach to flexible and accurate credit scoring modelling for Small and Large Enterprises using a database from a Greek Credit Bureau.

The originality and one of the main contributions of the proposed modelling methodology lies in the fact that we effectively blend financial features together with credit behavioural characteristics that have never been considered before. Furthermore, an algorithmic procedure that has been proposed and implemented into the methodology constitutes yet, another contribution since it is responsive to the need for dimension reduction, an issue frequently encountered in practice, especially in problems classified as falling into the area of BDA. For this, we rely on modern regularization and classification methods which ensure the construction of flexible yet, reliable credit scoring models. To the best of our knowledge, this is the first time that the combination of the above multivariate techniques is being used and implemented effectively, into credit scoring modelling.

As was mentioned earlier, among the advantages and contributions of the proposed methodology one could mention the originality in using a blend of financial characteristics and data related to the credit behaviour when authorities and institutions tend to rely almost solely on the former. The method is also appealing due to the use of popular logistic regression analysis. Indeed, it should be noted that after the 2-level dimension reduction procedure we choose to use standard logistic regression instead of other complex methods because it has proved its efficiency over the

years and is easily explained. Finally, the proposed methodology is responsive to the need for dimension reduction for the construction of a flexible yet reliable credit scoring model for purposes related to both prediction and description, with both financial and credit behaviour statistically significant variables.

One of the distinct advantages and contributions of this work lies in the character of the data used. Indeed, our findings clearly show the importance of using credit behavioural variables since a number of such variables have been found to play a key role in building credit scoring models both for Small and Large Enterprises. More specifically, in the final model for the Small Businesses, each PCA variable depends on 6 credit behavioural variables (out of a total of 15 variables) while for the Large Enterprises final model each PCA variable depends on 10 credit behavioural variables (out of a total of 18 variables). The use of such credit behavioural variables is undoubtedly one of the innovative findings of this work if one takes into consideration that countries and institutions rely almost solely for modelling purposes, on classical financial variables. Furthermore, the constant need for flexible yet accurate and reliable modelling approaches makes the proposed algorithmic procedure for dimension reduction, a valuable tool in the hands of researchers and practitioners. Indeed, it is also noteworthy that the proposed methodology provides among others, insurers, financial planners, and lenders with an automated reliable financial tool for evaluating creditworthiness according to a few statistically significant financial as well as credit behavioural variables and at the same time making credit decisions faster and fairer while offering to borrowers increased lending opportunities. In conclusion, the practical implications of these methodologies involve the construction of binary classification credit-scoring models based on Enterprises' data magnitude and peculiarities.

The dimension reduction modelling proposed in the present chapter may be extended and applied in fiscal debt credit scoring modelling. The significance of the prediction arising from fiscal debt rating agencies is another direction of possible extensions. Finally, the importance of predictions concerning both credit risk rating and fiscal debt rating models may be tested with the development of Support Vector Machines based on multiple kernels in conjunction with other approaches.

Chapter 9

Future Research

This dissertation provided the framework around some DRT that have been developed in an attempt to minimize the extent of the multicollinearity issue and at the same time reduce the dimensionality of a dataset. The manuscript utilized various multivariate analysis tools with the aim to study, analyse, compare and improve existing techniques and introduce new ones for handling multicollinearity and reducing the dimensionality of the resulted model. This Chapter is dedicated to the future expansion of the aforementioned works.

Chapter 4 concerned the modelling of PPE of various European countries. The identification, collection, and analysis of variables, which, either short-term or long-term, may have an impact on the shaping of the response variable was held. A combination of unsupervised DRT was implemented to obtain the optimal set of variables for the modelling of PPE. The analysis focused on 20 European countries for which a set of 20 possible explanatory variables for the period 2001–2015, were used. The model developed provides, with a minimum average error of fewer than 6‰ for each time period, accurate results for the PPE.

Using the above results as a first step, it is possible, depending on the data available, to develop in the immediate future, an evolved time series model that would be capable of predicting the Expenditures for 10-15 future years from the base year. The forecasting model could be used by any state that wishes to predict future Pension Expenditures based on its economy. This calculation primarily serves either as an estimate by itself or as a confirmation technique for the calculation of Expenditures made by other means.

Chapter 5 proposed the FS-PLS, a PLS-based method that acts as a feature selection and feature extraction DRT simultaneously, in linear regression tasks. In such a manner, we are able to remove the uninformative variables and obtain better or same results as the classical PLS regression but with a simpler structure both in univariate and multivariate scenarios. Concerning the possible future expansion of this work, an Elastic net-based FS-PLS is under consideration, for further reduction of the data.

We aim to investigate the cooperative effects of these two techniques on high-dimensional multicollinear data in order to make a projection on a low-dimensional space and thus construct less simplex and more interpretable linear regression models of high predictive accuracy with a penalized set of predictors.

Chapter 6 proposed and investigated a robust and easily interpretable methodology, termed EIC, capable of capturing multicollinearity rather accurately and effectively and thus providing a proper model assessment.

However, as mentioned in the dissertation, EIC is able to locate specific interdependency patterns, our aim so to manage to make it feasible for other dependency patterns. More precisely, the use of more complex patterns can be applied for data

coming from different distributions. Furthermore, the use of L_0 penalty can play a key role and possible be the answer to a more versatile EIC formula.

Chapter 7 attempted to locate and analyse via multivariate analysis techniques, highly correlated variables which were interrelated with the GDP and therefore are affecting either a short-term or a long-term shaping.

Taking that under consideration, it is possible to attempt to explore how different model selection criteria react or are able to make the right feature selection when multicollinearity is of a different magnitude. Through this process, one could be able to identify the criterion which is better adjusted and finally succeeds in choosing the optimal model when the variables involved are highly correlated.

The objective of Chapter 8 was the proposal of an innovative approach to flexible and accurate credit scoring modelling with the use of not only financial but also credit behavioural characteristics based on a multi-step DRT procedure. The resulting DRT-based modelling proposed in the present manuscript may be extended and applied to fiscal debt credit scoring modelling. The significance of the prediction arising from fiscal debt rating agencies is another direction of possible extensions.

Finally, the importance of predictions concerning both credit risk rating and fiscal debt rating models may be tested with the development of Support Vector Machines-based on multiple kernels in conjunction with other approaches. Additionally, the produced model can be applied for other countries and a comparison of it's effectiveness between Greece and other countries can be explored.

Bibliography

[1] Akaike, A. (1974). A New Look at the Statistical Model Identification. IEEE Transactions on Automatic Control, 19(6), 716–723.

[2] Akhter, Y., Mahsin, M. and Mohaimin, M.Z. (2012). An application of factor analysis on gross domestic product data of Bangladesh. Bangladesh e-Journal of Sociology, 9(1), 6 – 18.

[3] Alin, A. (2010). Multicollinearity. Wiley Interdisciplinary Reviews: Computational Statistics, 2(3), 370 – 374.

[4] Anderson, D. (2009). An Introduction to Multivariate Statistical Analysis. John Wiley and Sons.

[5] Artemiou, A. and Li, B. (2009). On Principal Components and Regression: A Statistical Explanation of a Natural Phenomenon. Statistica Sinica, 19, 1557 – 1565.

[6] Artemiou, A. and Li, B. (2013). Predictive Power of Principal Components for Single-Index Model and Sufficient Dimension Reduction. Journal of Multivariate Analysis, 119, 176 – 184.

[7] Bai, A., Hira, S. and Deshpande, P.S. (2015). An application of factor analysis in the evaluation of country economic rank. Procedia Computer Science, 54, 311 – 317.

[8] Barnes, P. (1987). The Analysis and Use of Financial Ratios: A Review Article. Journal of Business Finance and Accounting, 14(4), 449 – 461.

[9] Barr, P. (2006). The Economics of Pensions. Oxford Review of Economic Policy, 22(1), 15 – 39.

[10] Basel Committee on Banking Supervision (2004). Basel II: International Convergence of Capital Measurement and Capital Standards: A Revised Framework.

[11] Basu, A., Harris, I.R., Hjort, N.L. and Jones, M.C. (1998). Robust and efficient estimation by minimising a density power divergence. Biometrika, 85(3), 549 – 559.

[12] Bayer, S. (2018). Combining Value-at-Risk Forecasts using Penalized Quantile Regressions. Econometrics and Statistics, 8, 56–77.

[13] Beale, E.M.L., Kendall, M.G. and Mann, D.W. (1967). The Discarding of Variables in Multivariate Analysis. Biometrika, 54 (3 and 4), 357–366.

[14] Bellotti, T. and Crook, J. (2009). Support Vector Machines for Credit Scoring and Discovery of Significant Features. Expert Systems with Applications, 36(2), 3302 – 3308.

[15] Belsley, D. (1991). A Guide to Using the Collinearity Diagnostics. Computer Science in Economics and Management, 4(1), 33–50.

[16] Blanchard, O. (2000). Macroeconomics. New Jersey: Prentice-Hall.

[17] Boguslauskas, V., Mileris, R. and Adlyte, R. (2011). The Selection of Financial Ratios as Independent Variables for Credit Risk Assessment. Economics and Management, 16, 1032 – 1038.

[18] Bonoli, G. (2003). Two Worlds of Pension Reform in Western Europe. Comparative Politics, 35(4), 399–416.

[19] Bonoli, G. and Shinkawa, T. (2005). Ageing and Pension Reform Around the World: Evidence from Eleven Countries. Edward Elgar.

[20] Boritz, J.E. and Kennedy, D.B. (1995). Effectiveness of Neural Network Types for Prediction of Business Failure. Expert Systems with Applications, 9(4), 503 – 512.

[21] Boulesteix, A.L. and Strimmer, K. (2006). Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. Briefings in Bioinformatics, 8(1), 32 – 44.

[22] Callen, T. (2008). What is gross domestic product?. Finance and Development, 45(4), 48 – 49.

[23] Carone, G, Eckefeldt, P., Giamboni, L. and Laine, V., Pamies Sumner, S. and Economic and Financial Affairs (2016). Pension Reforms in the EU since the Early 2000's: Achievements and Challenges Ahead. European Economy Discussion Paper 042.

[24] Cattell, R.B. (1966). The Scree Test for the Number of Factors. Multivariate Behavioral Research, 4(1), 245–276.

[25] Cavanaugh, J.E. (2004). Criteria for Linear Model Selection Based on Kullback's Symmetric Divergence. Australian and New Zealand Journal of Statistics, 46(2), 257 – 274.

[26] Cheever, E. (2020). Course material. Swarthmore College, Department of Engineering.

[27] Chen, C., Ribeiro, B., Vieira, A.S. and Neves, J.C. (2011). A Genetic Algorithm-Based Approach to Cost-Sensitive Bankruptcy Prediction. Expert Systems with Applications, 38(10), 12939 – 12945.

[28] Corn dataset (2022). `http://software.eigenvector.com/data/index.html`.

[29] De Jong, S. (1993). SIMPLS: An alternative approach to partial least squares regression. Chemometrics and Intelligent Laboratory Systems, 18(3), 251 – 263.

[30] de La Fuente, A. (2015). A Simple Model of Aggregate Pension Expenditure. Hacienda Pública Española / Review of Public Economics, 212(1), 13–50.

[31] Diamonds, P. (2001). Towards an Optimal Social Security Design. CeRP Working Paper.

[32] Dumitrescu, B.A., Dedu, V. and Enciu A. (2009). The Correlation Between Unemployment and Real GDP Growth. A Study Case on Romania. Annals of Faculty of Economics, 2, 317 – 22.

[33] Eisenbeis, R.A. (1978). Problems in Applying Discriminant Analysis in Credit Scoring Models. Staff Studies 94, Board of Governors of the Federal Reserve System (U.S.).

[34] Eurostat database (2019). `https://ec.europa.eu/eurostat/data/database`.

[35] Faraway, J.J. (2002). Practical Regression and Anova using R. `https://julianfaraway.github.io/faraway/PRA/`.

[36] Farrar, D.E. and Glauber, R.R. (1967). Multicollinearity in Regression Analysis: The Problem Revisited. The Review of Economics and Statistics, 49, 92 – 107.

[37] Farrell, D. (2017). Risky Choices: Simulating Public Pension Funding Stress with Realistic Shocks. Hutchins Center Working Paper.

[38] Fisher, R. (1918). The Correlation Between Relatives on the Supposition of Mendelian Inheritance. Philosophical Transactions of the Royal Society of Edinburgh, 52, 399 – 433.

[39] Fisher, R. (1921). On the "Probable Error" of a Coefficient of Correlation Deduced from a Small Sample. Metron, 1.

[40] Franco, D.S., Marino M.R. and Zotteri, S. (2006). Pension Expenditure Projections, Pension Liabilities and European Union Fiscal Rules. SSRN.

[41] Fried, J. and Howitt, P. (1983). The Effects of Inflation on Real Interest Rates. The American Economic Review, 73, 968 – 980.

[42] Galton, F. (1886). Regression Towards Mediocrity in Hereditary Stature. The Journal of the Anthropological Institute of Great Britain and Ireland, 15, 246 – 263.

[43] Garcia-Zamor, J.C. (2018). The European Migrant and Refugee Crisis. In: Ethical Dilemmas of Migration. PAGG Springer, 29–44.

[44] Geary, R.C. and Leser, C.E.V. (1968). Significance Tests in Multiple Regression. The American Statistician, 22, 20–21.

[45] Giannouli, P. and Kountzakis, C. (2019). Towards an Improved Credit Scoring System: The Greek Case. International Journal of Financial Engineering and Risk Management, 3(1), 19 – 31.

[46] Gibson, H.D., Palivos, T. and Tavlas, G.S. (2014). The Crisis in the Euro Area: An Analytic Overview. Journal of Macroeconomics, 39(2), 233–239.

[47] Greene, W. (2002). Econometric Analysis, 5th ed.. Hoboken: Prentice Hall.

[48] Gujarati D.N. and Porter, D.C. (2008). Basic Econometrics, 5th ed.. New York: Mc-Graw Hill.

[49] Guttman, L. (1954). Some Necessary Conditions for Common-Factor Analysis. Psychometrika, 19, 149–161.

[50] Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, 3, 1157 – 1182.

[51] Hair, J.F., Black, W.C., Badin, B.J. and Anderson, R.E. (2010). Advanced Diagnostics for Multiple Regression: A Supplement to Multivariate Data Analysis. Pearson Prentice Hall Publishing.

[52] Halkos, G. and Tsilika, K. (2018). Programming Correlation Criteria with free CAS Software. Computational Economics, 52, 299 – 311.

[53] Hall, M. (1999). Correlation-based feature selection for machine learning. The University of Waikato,Hamilton, New Zealand.

[54] Hardy, W. and Adrian, J.L. (1985). A Linear Programming Alternative to Discriminant Analysis in Credit Scoring. Agribusiness, 1(4), 285 – 292.

[55] Hastie, T., Tibshirani, R. and Friedman, J. (2001). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer.

[56] Helland, I. (1988). On the structure of partial least squares regression. Communications in Statistics - Simulation and Computation, 17(2), 581 – 607.

[57] Hickman, J. (1968). Funding Theories for Social Insurance. Casualty Actuarial Society, 55, 303-311.

[58] Holzmann, R. (2009). Aging Population, Pension Funds, and Financial Markets: Regional Perspectives and Global Challenges for Central, Eastern, and Southern Europe. Directions in development – finance, World Bank.

[59] Homburg, S. (2000). Compulsory Savings in the Welfare State. Journal of Public Economics, 77, 233–239.

[60] Honeine, P., Mouzoun, S. and Eltabach, M. (2018). Neighbor Retrieval Visualizer for Monitoring Lifting Cranes. Advances in Condition Monitoring of Machinery in Non-Stationary Operations: Proc. 6th International Conference on Condition Monitoring of Machinery in Non-stationary Operations, Santander, Spain.

[61] Höskuldsson, A. (1988). PLS regression methods. Journal of Chemometrics , 2(3), 211 – 228.

[62] Hotelling, H. (1933). Analysis of a Complex of Statistical Variables into Principal Components. Journal of Educational Psychology, 24, 417—441 and 498—520.

[63] Hotelling, H. (1936). Relations between Two Sets of Variates. Biometrika, 28 3/4, 321-–377.

[64] Imdadullah, M., Aslam, M. and Altaf, S. (2016). mctest: An R Package for Detection of Collinearity among Regressors. The R Journal, 8, 495 – 505.

[65] International Business Machines Corporation (2021) What is Big Data Analytics?. https://www.ibm.com/analytics/big-data-analytics.

[66] Jolliffe, I. (1972). Discarding Variables in a Principal Component Analysis. I: Artificial Data. Journal of the Royal Statistical Society. Series C (Applied Statistics), 21, 160–173

[67] Jolliffe, I. (2002). Principal Components Analysis. Springer, 2nd ed..

[68] Kaiser, H. (1960). The Application of Electronic Computers to Factor Analysis. Educational and Psychological Measurement, 20, 141–151.

[69] Kalivas, J. (1997). Two data sets of near infrared spectra. Chemometrics and Intelligent Laboratory Systems, 37(2), 255 – 259.

[70] Karam, P.D., Muir, D., Pereira, J. and Tuladhar, A. (2010). Macroeconomic Effects of Public Pension Reforms. International Monetary Fund (10/297).

[71] Kendall, M. (1957). A Course in Multivariate Analysis. New York:Hafner Pub. Co..

[72] Klein, L. (1962). An Introduction to Econometrics. Englewood Cliffs: Prentice Hall.

[73] Knoema database (2022). `https://knoema.com`.

[74] Kondo, M., Mizuno, O. and Choi, E.H. (2018). Causal-effect analysis using Bayesian LiNGAM comparing with correlation analysis in function point metrics and effort. International Journal of Mathematical, Engineering and Management Sciences, 3(2), 90 – 112.

[75] Kovács, P., Petres,T. and Tóth, L. (2005). New Measure of Multicollinearity in Linear Regression Models. International Statistical Review/Revue Internationale de Statistique , 73, 405 – 12.

[76] Kozak M. and Piepho, H.P. (2018). What's normal anyway? Residual plots are more telling than significance tests when checking ANOVA assumptions. Journal of Agronomy and Crop Science, 204, 86 – 98.

[77] Kullback, S. and Leibler, R.A. (1951). On information and sufficiency. The Annals of Mathematical Statistics, 22(1), 79 – 86.

[78] Kumar, U. (2005). Comparison of Neural Networks and Regression Analysis: A New Insight. Expert Systems with Applications, 29(2), 424 – 430.

[79] Laboratory of Statistics and Data Analysis (2022). `http://actuarweb.aegean.gr/labstada/publications.html`.

[80] Lachowska, M. and Myck, M. (2018). The Effect of Public Pension Wealth on Saving and Expenditure. American Economic Journal: Economic Policy, 10, 284 – 308.

[81] Largey, J. (1996). F- and T-Tests in Multiple Regression: The Possibility of 'Conflicting' Outcomes. Journal of the Royal Statistical Society. Series D (The Statisician), 45, 105—109.

[82] Lendaro, A. (2016). A European Migrant Crisis? Some Thoughts on Mediterranean Borders. Studies in Ethnicity and Nationalism, 16, 148 – 157.

[83] Li, B. (2018). Sufficient dimension reduction: methods and applications with R. C. Chapman and Hall/CRC.New York.

[84] Li, B., Morris, J. and Martin, E.B. (2002). Model selection for partial least squares regression. Chemometrics and Intelligent Laboratory Systems, 64(1), 79 – 89.

[85] Li, G.Z. and Zeng, X.Q. (2009). Feature Selection for Partial Least Square Based Dimension Reduction. Foundations of Computational Intelligence, 5, 3 – 37.

[86] Lindgren, F., Geladi, P. and Wold, S. (1993). The kernel algorithm for PLS. Journal of Chemometrics, 7(1), 45 – 59.

[87] Lindner, T., Puck, J. and Verbeke, A. (2020). Misconceptions about Multicollinearity in International Business Research: Identification, Consequences, and Remedies. Journal of International Business Studies, 51, 283 – 298.

[88] Liu, W. and Li, Q. (2017). An Efficient Elastic Net with Regression Coefficients Method for Variable Selection of Spectrum Data. PLOS ONE , 12: e0171122.

[89] Mantalos, P., Mattheou, K. and Karagrigoriou, A. (2010). An improved divergence information criterion for the determination of the order of an AR process. Communications in Statistics—Simulation and Computation, 39(5), 865 –879.

[90] Marcinkiewicz, E. and Chybalski, F. (2014). How to Measure and Compare Pension Expenditures in Cross-Country Analyses? Some Methodological Remarks. International Journal of Business and Management, 2, 43 – 59.

[91] Marcinkiewicz, E. and Chybalski, F. (2016). A New Proposal of Pension Regimes Typology: Empirical Analysis of the OECD Countries. ENRSP Conference, Poland.

[92] Marinescu, D.E. and Manafi, I. (2017). The Effects of International Migration on the Pension Systems in Europe. Tér és Társadalom, 31(4), 40–52.

[93] Martens H. and Naes, T. (1989). Multivariate Calibration. New York: Wiley.

[94] Mattheou, K., Lee, S. and Karagrigoriou, A. (2009). A model selection criterion based on the BHHJ measure of divergence. Journal of Statistical Planning and Inference, 139(2), 228 – 235.

[95] Mauri, M., Elli, T., Caviglia, G., Uboldi, G. and Azzi, M. (2017). RAWGraphs: A Visualisation Platform to Create Open Outputs.In Proceedings of the 12th Biannual Conference on Italian SIGCHI Chapter. New York: Association for Computing Machinery.

[96] Mavri, M. and Ioannou, G. (2004). An Empirical Study for Credit Card Approvals in the Greek Banking Sector. Operational Research. An International Journal, 4(1), 29 – 44.

[97] Mavri, M., Angelis, V., Ioannou, G., Gaki, E. and Koufodontis, I. (2008). A Two-Stage Dynamic Credit Scoring Model, Based on Customers Profile and Time Horizon. Journal of Financial Services Marketing, 13, 17 – 27.

[98] McCullagh, P. and Nelder, J.A. (1989). Generalized Linear Models. Chapman Hall.

[99] Müller, K. (2001). The Political Economy of Pension Reform in Eastern Europe. International Social Security Review, 54(2-3), 57-79.

[100] Muto, I., Oda, T. and Sudo, N. (2016). Macroeconomic Impact of Population Aging in Japan: A Perspective from an Overlapping Generations Model. IMF Economic Review, 64(3), 408-442.

[101] Myers, J.H. and Forgy, E.W. (1963). The Development of Numerical Credit Evaluation Systems. Journal of the American Statistical Association, 58(303), 799 – 806

[102] Ntotsis K. and Karagrigoriou, A. (2021). The Impact of Multicollinearity on Big Data Multivariate Analysis Modeling. In Applied Modeling Techniques and Data Analysis 1: Computational Data Analysis Methods and Tools. London: iSTE Ltd., 187 – 202.

[103] Ntotsis, K., Papamichail, M., Hatzopoulos, P. and Karagrigoriou, A. (2019). On the Modelling of Pension Expenditures in Europe. Communications in Statistics: Case Studies, Data Analysis and Applications, 6(1), 50 – 68.

[104] Ntotsis, K., Kalligeris, E.N. and Karagrigoriou, A. (2020). A Comparative Study of Multivariate Analysis Techniques for Highly Correlated Variable Identification and Management. International Journal of Mathematical, Engineering and Management Sciences, 5(1), 45 – 55.

[105] O'Rourke, N., Hatcher, L., Stepanski, E.J. and SAS Institute, Inc. (2005). A Step-by-Step Approach to Using SAS for Univariate and Multivariate Statistics, 2nd ed. Cary, NC: SAS Institute.

[106] Oner, C. (2020). Unemployment: The Curse of Joblessness. International Monetary Fund.

[107] Organisation for Economic Co-operation and Development database (2022). `https://data.oecd.org`.

[108] Organisation for Economic Co-operation and Development (2019). Definition for gross domestic product. `https://data.oecd.org/gdp/gross-domestic-product-gdp.htm`.

[109] Organisation for Economic Co-operation and Development (2020). Pension Expenditures definition. `https://data.oecd.org/socialexp/pension-spending.htm`.

[110] Organisation for Economic Co-operation and Development (2021). OECD Main Economic Indicators (MEI). `https://www.oecd.org/sdd/oecdmaineconomicindicatorsmei.htm`.

[111] Paleologo, G., Elisseeff, A. and Antonini, G. (2010). Subagging for Credit Scoring Models. European Journal of Operational Research, 201(2), 490 – 499.

[112] Pearson, K. (1901). On Lines and Planes of Closest Fit to Systems of Points in Space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2, 559 – 572.

[113] Pedregosa F., Varoquaux, G., Gramfort, A., Michel, V. and Thirion, B. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825 – 2830. *https://scikit-learn.org/stable/autoexamples/crossdecomposition/plotpcrvspls.html*

[114] Plamondon, P., Drouin, A., Binet, G., Cichon, M., McGillivray, W.R., Bedard, M. and Perez-Montas, H. (2003). Actuarial Practice in Social Security. Geneva: International Labor Office.

[115] Perez-Melo, S. and Golam Kibria, B.M. (2020). On Some Test Statistics for Testing the Regression Coefficients in Presence of Multicollinearity: A Simulation Study. Stats, 3, 40 – 55.

[116] Pagès, J. (2015). Multiple Factor Analysis by Example Using R. New York: Taylor and Francis Group/CRC.

[117] Reijer, A. (2005). Forecasting Dutch GDP using large scale factor models. DNB Working Papers 028, Netherlands Central Bank, Research Department.

[118] Rosipal R. and Kramer, N. (2005). Overview and Recent Advances in Partial Least Squares. Saunders C., Grobelnik M., Gunn S., Shawe-Taylor J. (eds) Subspace, Latent Structure and Feature Selection. SLSFS. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 3940.

[119] Samuelson, P. (1958). An Exact Consumption-Loan Model of Interest with or without the Social Contrivance of Money. The Journal of Political Economy, 66(6), 467 – 82.

[120] Scheffe, H. (1999). The Analysis of Variance. John Wiley and Sons.

[121] Schneider, O. (2005). Pension Reform: How Macroeconomics May Help Microeconomics - The Czech Case. SSRN Electronic Journal.

[122] Schumacher, C. (2007). Forecasting German GDP using alternative factor models based on large datasets. Journal of Forecasting, 26(4), 271 – 302.

[123] Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics, 6(2), 461 – 464.

[124] Shang, J. and Cavanaugh, J.E. (2008). Bootstrap Variants of the Akaike Information Criterion for Mixed Model Selection. Computational Statistics and Data Analysis, 52(4), 2004–2021.

[125] Sheather, S. (2009). A Modern Approach to Regression with R. Springer Science and Business Media.

[126] Siddiqi, N. (2006). Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring. Hoboken, NJ: John Wiley and Sons, Inc..

[127] Siddiqi, N. (2016). Intelligent Credit Scoring: Building and Implementing Better Credit Risk Score-cards, 2nd ed.. New Jersey: John Wiley and Sons.

[128] Silvey, S. (1969). Multicollinearity and Imprecise Estimation. Journal of the Royal Statistical Society. Series B, 31, 539 – 52.

[129] Smallman, L., Artemiou, A. and Morgan, J. (2018). Sparse Generalised Principal Component Analysis. Pattern Recognition, 83, 443 – 455.

[130] Theil, H. (1971). Principles of Econometrics. New York: John Wiley and Sons.

[131] Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B, 58, 267 – 88.

[132] Tikhonov, A. (1943). On the Stability of Inverse Problems. Doklady Akademii Nauk SSSR, 39, 195 – 98.

[133] Tikhonov, A. (1963). Solution of Incorrectly Formulated Problems and the Regularization Method. Soviet Mathematics, 4, 1035 – 38.

[134] Toma, A. (2014). Model Selection Criteria Using Divergences. Entropy, 16(5), 2686 – 2689.

[135] Trading Economics (2011). Main Indicators. `https://tradingeconomics.com/indicators`.

[136] Trygg, J. and Wold, S. (2002). Orthogonal projections to latent structures (O-PLS). Journal of Chemometrics, 16(3), 119 – 128.

[137] Ueki, M and Kawasaki, Y. (2013). Multiple Choice from Competing Regression Models under Multicollinearity based on Standardized Update. Computational Statistics and Data Analysis, 63, 31 – 41.

[138] Ullah M.I., Aslam, M., Altaf, S. and Ahmed, M. (2019). Some New Diagnostics of Multicollinearity in Linear Regression Model. Sains Malaysiana, 48, 2051 – 60.

[139] Varmuza, K. and Filzmoser, P. (2009). Introduction to Multivariate Statistical Analysis in Chemometrics. CRC Press.

[140] Wang, L. and Alexander, C.A. (2019). Big data analytics in healthcare systems. International Journal of Mathematical, Engineering and Management Sciences, 4(1), 17-26.

[141] Wasserman, L. (2004). All of Statistics: A Course in Statistical Inference. Springer, New York.

[142] Wehrens, H. (2011). Chemometrics with R: multivariate data analysis in the natural sciences and life sciences. New York, NY: Springer.

[143] Weisburd, D. and Britt, C. (2013). Statistics in Criminal Justice. Berlin/Heidelberg: Springer Science and Business Media.

[144] Wold S., Sjöström M. and Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems, 58(2), 109 – 130.

[145] Wold, H. (1975). Path models with latent variables: The NIPALS approach. H.M. Blalock et al., editor, Quantitative Sociology: International perspectives on mathematical and statistical model building, Academic Press., 307 – 357.

[146] Wooldridge, J. (2014). Introduction to econometrics: Europe, Middle East and Africa Edition. Boston: Cengage Learning.

[147] World Bank Open Data (2021).

[148] Yamashita, T., Yamashita, K. and Kamimura, R. (2007). A Stepwise AIC Method for Variable Selection in Linear Regression. Communications in Statistics – Theory and Methods, 36(13), 2395 – 2403.

[149] Yeniay, O. and Goktas, A. (2002). A comparison of partial least squares regression with other prediction methods. Hacettepe Journal of Mathematics and Statistics, 31.

[150] Yoo, W., Mayberry, R., Bae, S., Singh, K., He, Q.P. and Lillard, J.W. (2014). A Study of Effects of Multicollinearity in the Multivariable Analysis. International Journal of Applied Science and Technology, 4(5), 9 – 19.

[151] Yu, L, Yue, W., Wang, S. and Lai, K. (2010). Support Vector Machine Based Multiagent Ensemble Learning for Credit Risk Evaluation. Expert Systems with Applications, 37(2), 1351 – 1360.

[152] Yue, L., Li, G., Lian, H. and Wan, X. (2019). Regression Adjustment for Treatment Effect with Multicollinearity in High Dimensions. Computational Statistics and Data Analysis, 134, 17 – 35.

[153] Zhang, Z. (2016). Variable Selection with Stepwise and Best Subset Approaches. Annals of Translational Medicine, 4(7), 136.

[154] Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B, 67(2), 301 – 320.

[155] Zou, H., Hastie, T. and Tibshirani, R. (2006). Sparse Principal Component Analysis. Journal of Computational and Graphical Statistics, 15, 265 – 86.

Appendix A

Linear Regression Analysis

This is an optional Chapter and can be considered as an introduction to statistical modelling, which takes place in many Chapters of this dissertation. For more, see Sheather (2009) and Wasserman (2004).

Analysis of Variance

ANOVA, or analysis of variance, is a statistical technique developed by Fisher (1918), (1921), where he introduced the terms variance and analysis of variance is used to compare two or more populations while different types of effects operate concurrently. Essentially, it is a procedure that determines whether those effects are significant, what their estimation is, and whether there are differences between population means (Scheffe (1999)).

An ANOVA table is formed as follows:

Source	Sum of Squares	DF	Mean Squared	F	Sig.
Explained	ESS	k-1	MSE = ESS / k-1	F = MSE / MSR	
Residuals	RSS	n-k	MSR = RSS / n-k		
Total	TSS	n-1			

Note: *In some cases, the symbols may differ. Explained Sum of Squares, ESS, can be seen as regression sum of squares, which is represented by RSS, and residual sum of squares, RSS, can be seen as error sum of squares, which is represented by ESS. To understand the procedure, the reader must focus on the meaning behind the symbolisation.*

ANOVA interpretation

The purpose of this procedure is to see if there are any differences between the methods mentioned above. The first step in accomplishing this is to create a hypothesis test. In statistics, hypothesis testing is a method of determining whether the results of a survey or experiment are valid and meaningful. The following are the components of a hypothesis test:

1. **Null and alternative hypotheses**.
   The null hypothesis, denoted by H_0, is always the accepted fact, while the alternative, denoted by H_α, is the one that is questionable and must be examined.

2. **Predetermined level of significance**
   The significance level is defined by the "tolerance" given by the experiment's conductor in the presence of error type I.

Note: *When conducting a statistical experiment, the risk of incorrect decision-making may occur. There are two types of errors that may happen.*

- Type I error: reject a true H_0
- Type II error: failing to reject a false H_0

Let us define,

$$\alpha = P(type\ I\ error)$$

and

$$\beta = P(type\ II\ error)$$

Then, one of the following options about the experiment decision occurred

Options	Fail to reject null hypothesis	Reject null hypothesis
H_0 **is true**	- with probability = 1-α	Type I error with probability = α
H_α **is false**	Type II error with probability = β	- with probability = 1-β

3. **Test statistic and critical zone of the test**
   It used when deciding whether or not the null hypothesis should be rejected. When $H0$ is true, it is a random quantity with a known distribution. It is based on the $Halpha$, the distribution of the test statistic, and the $alpha$ significance level. The $Halpha$ value defines the formation of the critical zone.
4. **Value of the test statistic**
   The test value is computed based on the sample values, and if it exists in the critical zone, the null hypothesis is rejected; otherwise, it is not rejected.
5. **The decision to reject or not the null hypothesis**
   The experiment's conductor decides whether or not to reject the null hypothesis.

Test Hypothesis

Following that, the AVONA table null hypothesis is used to determine whether there are any differences between the means. The null and alternatives are written as follows:

$$H_0 : \mu_1 = \mu_2 .... = \mu_n$$

$$H_\alpha : At\ least\ one\ \mu_i\ differs\ from\ the\ others$$

where i=1,2,...,n.

ESS

Explained Sum of Squares, denoted as ESS, expresses the variability between sampling means as the sum of the squares of the distances of each medium from the total mean.

$$ESS = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

RSS

Residual Sum of Squares, denoted as RSS, expresses the variability between the sampling means, which is measured as the sum of the squares of the distances of each medium with the total mean.

$$RSS = \sum_{i=1}^{n} \hat{\varepsilon}_i^2$$

TSS

Total Sum of Squares, denoted by *TSS*, expresses the overall variability of the observations

$$\begin{aligned}
\sum_{i=1}^{n}(Y_i - \overline{Y})^2 &= \sum_{i=1}^{n}(Y_i - \overline{Y} + \hat{Y}_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}((\hat{Y}_i - \bar{Y}) + \underbrace{(Y_i - \hat{Y}_i)}_{\hat{\varepsilon}_i})^2 \\
&= \sum_{i=1}^{n}((\hat{Y}_i - \bar{Y})^2 + 2\hat{\varepsilon}_i(\hat{Y}_i - \bar{Y}) + \hat{\varepsilon}_i^2) \\
&= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}\hat{\varepsilon}_i^2 + 2\sum_{i=1}^{n}\hat{\varepsilon}_i(\hat{Y}_i - \bar{Y}) \\
&= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}\hat{\varepsilon}_i^2 + 2\sum_{i=1}^{n}\hat{\varepsilon}_i(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip} - \overline{Y}) \\
&= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}\hat{\varepsilon}_i^2 + 2(\hat{\beta}_0 - \overline{Y})\underbrace{\sum_{i=1}^{n}\hat{\varepsilon}_i}_{0} \\
&\quad + 2\hat{\beta}_1\underbrace{\sum_{i=1}^{n}\hat{\varepsilon}_i x_{i1}}_{0} + \cdots + 2\hat{\beta}_p\underbrace{\sum_{i=1}^{n}\hat{\varepsilon}_i x_{ip}}_{0} \\
&= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \text{ESS} + \text{RSS}
\end{aligned}$$

DF

Degrees of freedom, denoted by *DF*, of an estimate is the number of independent pieces of information that went into calculating the estimate and in particular:

- $\mathbf{k - 1}$ are the *DF* of the divergence from the H_0
- $\mathbf{n - k}$ are the *DF* of residuals

- $\mathbf{n-1}$ are the total *DF*

MSE

Explained mean square, denoted by *MSR*, is defined by the error between the sample.

$$\text{MSE} = \frac{1}{k-1}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

MSR

Mean square of residuals, denoted by *MSE*, is defined by the error within the sample.

$$\text{MSR} = \frac{1}{n-k}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

F-test

As previously stated, the F-test is the ratio of variation between samples and variation within samples. This is the test that ANOVA employs to validate the test hypothesis.

If the test fails to reject the H_0, then,

$$F^* \sim F_{(k-1),(n-k)}$$

Otherwise,

$$F^* > F_{(k-1),(n-k),\alpha}$$

Sig.

The significant value of the F-test is represented by Sig. in the ANOVA table. Is most commonly known as the p-value, which is defined as the probability of observing a random price of test statistics that is equal to or more extreme than the observed one in terms of H_0, given that H_0 is rejected. In essence, it is the lowest significance level at which the H_0 is rejected.

Decision cases about the hypothesis test

- If $p\text{-}value < \alpha$, then the H_0 is rejected
- If $p\text{-}value > \alpha$, then the H_0 is not rejected
- If $p\text{-}value = \alpha$, then no decision about the rejection or not of the H_0 can be made

Regression Analysis

In statistics, the term regression was first introduced and used by Galton (1886) during an experiment in which he introduced the term regression to mediocrity. Sheather (2009) defines regression analysis as "the study of the dependence of one variable, the dependent variable, on one or more other variables, the explanatory

variables, with a view to estimating and/or predicting the (population) mean or average value of the former in terms of the known or fixed (in repeated sampling) values of the latter.'.

Linear Regression

This type of regression is used when there is a linear relationship between the response variable textit(dependent and the explanatory variable(s) (independent).

Simple linear regression

This type of regression takes place when only one independent variable exists. The model formation is:

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

where, α and β are called coefficients of regression. More specific α is the value of the dependent variable Y corresponding to the value X=0 of the dependent variable and β is the variation of the dependent variable corresponding to a unit change of X. Finally, ε_i is the error term that represents the deviation of the observed value from the true value of the quantity of interest.

Multiple linear regression

This type of regression takes place when more than one independent variable exists. The model formation is:

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + ...\beta_p X_{ip} + \varepsilon_i$$

A typical linear regression analysis consists of the following:

- Estimates of coefficients
- Standard Error (SE)
- T- statistics
- P-value of t-statistics
- Number of observations
- Error degrees of freedom
- Root Mean Squared Error
- R-square
- Adjusted R-square
- F-statistics
- P-value of F-statistics

Interpretation of Linear Regression Elements

The Interpretation of the elements which constitute the *Linear Regression Analysis*, with X_1,...,X_p the number of the existing independent variables, it follows.

- **Estimate**
  It displays the values for the regression coefficients for predicting the dependent variable from the independent variable, i.e., the values of α, β_1, ..., β_p.

- **SE**
  SE displays the standard error of each coefficient.

- **T-statistics**
  It displays the t-statistic values, namely the values of the Student's t-test. A t-test is commonly used to determine whether a regression coefficient is significant; i.e., whether it differs or not from zero.
  In other words, the null hypothesis of this test is used to decide whether each variable is statistically significant. The null and alternative are of the form:

$$H_0 : \beta_i = 0$$

$$H_\alpha : \beta_i \neq 0$$

  where *i=1,2,...,p*

- **P-value**
  This column displays the 2-tailed *p-values* associated with the t-test and are used to determine whether a given coefficient is significantly different from zero.

- **Number of observations**
  The number of observations is the size of the sample.

- **Error degrees of freedom**
  As mentioned before, in *ANOVA* interpretation, the degrees of freedom of an estimate is the number of independent pieces of information that enter into the estimate calculation.

- **Root Mean Squared Error**
  Root Mean Squared Error, denoted by *(RMSE)*, is defined by the standard deviation of the variance. The *RMSE* of an estimator $\hat{\theta}$ is defined by:

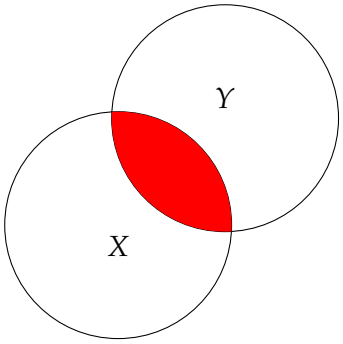
$$\mathrm{RMSE}(\hat{\theta}) = \sqrt{\mathrm{MSE}(\hat{\theta})}$$

- **R-square**
  R-square, denoted by R^2, represents the percentage of the total variability of

the dependent variable interpreted on the basis of the regression model

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{SST}$$

R^2 can be also seen via a Venn diagram. The one below displays a simple linear regression model with one dependent and one independent variable. R^2 is the two-circle intersection*, which shows the extent to which the variation of Y is interpreted by the variation of X



- **Adjusted R-Square**
  Adjusted R-Square, denoted by R^2_{adj}, is used to decide about the usefulness of the independent variables in the model. The addition of a useless variable to the model, will cause decrease to the adjusted R-square, while the addition of a useful variable, will cause increase, but will never exceeds the R^2.

- **F-statistics**
  F-test and the associated p-value have been thoroughly analyzed in ANOVA interpretation.

Regression Assumptions

Analysis of variance as well as regression analysis answers some "questions" about the given data set that being analyzed. It should be noted though that those answers are not always trustworthy, or as in the field of statistics referred as statistically significant. To come to the conclusion that the model is significant some assuption must be fulfilled. Those assumptions are called linear regression assumptions and are based on the residuals ε_i of the model. Depending on the formation (i.e., simple, multiple) those assumptions might differ from case to case, but the following four are the most important and must be satisfied in every model formations.

- Normality
- Independence
- Homoscedasticity
- Linearity between Y and X_i

Normality

The errors must follow a normal distribution with zero mean and σ^2 variance, symbolized by:

$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

Most common tests that check this assumption are the *Lilliefors* test for normality, which is an improvement of the *Kolmogorov-Smirnov* test as well as the *Shapiro-Wilk* normality test.

$$H_0 : \textit{The residuals come from a standard normal distribution}.$$

$$H_\alpha : \textit{The residuals do not come from such a distribution}.$$

When it comes to real data sets, usually this assumption is not satisfied. For that reason some transformations have been proposed, which correct this problem. These transformations are mostly based on functions of the dependent variable with the Logarithm, Root, and Box and Cox transformations being the most popular ones.

Note: *The following theorem establishes the condition for residual independence.*
If the residuals are normally distributed and uncorrelated, then they are independent. The inverse of the theorem is not true (Gujarati and Porter (2008)).

Independence

The residuals must be independent. There are two methods that are commonly used to decide on independence. The first is with the execution of the Runs test for independence, which states the following hypotheses:

$$H_0 : \textit{The values of the residuals come in random order}.$$

$$H_\alpha : \textit{The values of the residuals do not come in random order}.$$

The second way to check independence is based on the previous theorem. After the normality has been checked, one can check the assumption of correlation via a Durbin-Watson test with the following hypotheses.

$$H_0 : \textit{The residuals are not autocorrelated}.$$

$$H_\alpha : \textit{The residuals are autocorrelated}.$$

Homoscedasticity

The errors must have the same finite variance. Levene's test for homoscedasticity is the most common test that checks this assumption, which states the following hypotheses:

$$H_0 : \textit{All population variances are equal}.$$

H_α : *At least one population variance differs from the others.*

Linearity

Linearity between Y and X_i can be seen through a scatter diagram.

Note: *The assumption of multicollinearity, or simply collinearity, must also be checked in multiple or multivariate regression.*

Graphical Assumptions Interpretation

In recent years, an increasing number of statisticians have begun to use graphical representations of their data to draw conclusions about various tests. This is not an exception when it comes to ANOVA assumptions. This point of view grows stronger by the day as new articles supporting this theory are published, with one of the most influential being an article by Kozak and Piepho (2018).

The graphs that can check the assumptions above are:

- **Normality**
  The symmetry plot of residuals can be used for the interpretation of normality. A symmetrical distribution of the residuals around their median suggests the existence of normal distribution.

- **Independence**
  The residuals versus lagged residuals plot can be used for the interpretation of correlation. A trend among the residuals indicates a possible correlation between them. If the residuals plots confirm the assumptions of correlations and normality, then the residuals are independent.

- **Homoscedasticity**
  The residuals versus the fitted values plot can be used for the interpretation of homoscedasticity. The increase in the variance as the fitted values increase suggests possible heteroscedasticity.

- **Linearity**
  Residuals versus every single one independent variable of the model.

Appendix B

Supplementary Material for Chapter 8: Multistep Dimension Reduction for Credit Scoring Modelling

This Appendix constitutes the supplementary material of the work entitled "Multilevel Dimension Reduction for Credit Scoring Modelling and Prediction: Empirical Evidence for Greece" and gives the definitions of the selected variables for both Small and Large Enterprises. The majority commentary was derived from Tiresias S.A., private online library.

Variables appearing only in Small Enterprises:

1. **Debt Equity Ratio** = Total Liabilities / Shareholder Equity. This ratio is used to evaluate a enterprise financial leverage.
2. **Return on Equity (ROE)** = Net Income/Average Shareholders' Equity. Roe is considered a measure of how effectively management is using a enterprise's assets to create profits.
3. **Working Capital Leverage** = Current Liabilities / Working Capital. Working capital leverage refers to the impact of level working capital on business's profitability. The working capital management should improve the productivity of investments in current assets and ultimately it will increase the return on capital employed.
4. **Total Assets Turnover Ratio** = Net Sales / Total Assets. This ratio measures a enterprise's ability to generate sales from its assets by comparing net sales with average total assets. It calculates net sales as a percentage of assets to show how many sales are generated from each dollar of enterprise assets.
5. **Return on Assets (ROA)** = Net Income / Total Assets. ROA is an indicator of how profitable a enterprise is relative to its total assets. ROA gives an idea to how efficient a business management is at using its assets to generate earnings.
6. **Result Curried Forward** = profits / damages.
7. **Profit Before Taxes Depreciation and Amortization Expense** = a profitability measure that looks at a enterprise's profit before the enterprise has to pay corporate income tax and depreciation and amortization expense.

Variables appearing only in Large Enterprises:

1. **Cash Ratio** = the ratio of a enterprise's total cash and cash equivalents to its current liabilities and signifies the enterprise's ability to pay short-term liabilities with its highest liquid assets.
2. **Current Assets to Total Liabilities** = measures the enterprise's ability to cover its total liabilities with its total current assets. This ratio is also used to estimate the liquidity of the enterprise by showing the enterprise can pay its creditors with its current assets if the business's assets ever had to be liquidated.
3. **Net Profit Margin** = net profit / revenue. This ratio is used to calculate the percentage of profit a business produces from its total revenue.
4. **Current Liabilities Turnover Ratio** = (short-term liabilities / net revenues from sales)* number of days in the period. This ratio indicates the number of days from the moment some liability arises to the moment it is paid.
5. **Fixed Assets to Equity** = fixed assets / equity. It measures the contribution of stockholders and the contribution of debt sources in the fixed assets of the enterprise.
6. **Working Capital Turnover Ratio** = net annual sales / average working capital. This ratio measures how efficiently a enterprise is using its working capital to support a given level of sales.
7. **Long-term Liabilities** = an obligation resulting from a previous event that is not due within one year of the date of the balance sheet.
8. **Total fixed Assets (net book value)** = Its formula is calculated by subtracting all accumulated depreciation and impairments from the total purchase price and improvement cost of all fixed assets reported on the balance sheet.
9. **Maximum Utilization- Not Revolving** = RCS Maximum percent credit utilization – Joint / Prime – Non-Revolving- SME – Updated in last 12 months.
10. **Total Current Balance** = RCS Total Current Balance – Joint / Prime – Open. When referring to a loan such as an auto loan or a mortgage, your current balance is the amount you currently still owe on the loan according to the date of your statement.

Variables appearing both Small and Large Enterprises:

1. **Total Liabilities** = the aggregate of all debts an individual or enterprise is liable for and can be calculated by summing all short-term and long-term liabilities.
2. **Short-term Liabilities** = a financial obligation that is to be paid within one year.
3. **Worst Payment Status Last Month vs Last 24 Months** = RCS Worst Payment Status – Joint / Prime – Last 1 Month vs. Last 24 Months.
4. **Worst Payment Status in Last 3 Months** = Worst Payment Status – SME – Joint/Prime – Last 3 Months.

5. **Maximum Number of Months Consecutive with over 100% Utilization in Last 24 Months** = RCS Maximum Number of Months Consecutive with over 100% of Percentage Credit Utilization in last 24 months - updated in last 12 months – Joint/Prime.

6. **Number of Occurrences with Delinquency 1+ in Last 12 Months** = RCS Number of Occurrences of Delinquency 1+ DPD – Joint/Prime – last 12 months.

7. **Current Balance / Delinquency to Current Balance** = RCS Ratio Current Balance / Delinquency to Current Balance – Joint / Prime – Open – updated in last 3 months.

8. **Maximum Number of Months Consecutive with over 100% Utilization in Last 6 Months** = RCS Maximum Number of Months Consecutive with over 100% of Percentage Credit Utilization in last 6 Months – Updated in last 3 months – Joint / Prime.