ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Σχολή Χρηματοοικονομικής και Στατιστικής

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

Σύγχρονες μέθοδοι ανάλυσης δεδομένων επιβίωσης υψηλής διάστασης: Επιλογή μεταβλητών και μοντέλα πρόβλεψης

Νικόλαος Σγουρός

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην Εφαρμοσμένη Στατιστική

Πειραιάς

Σεπτέμβριος 2026

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Σχολή Χρηματοοικονομικής και Στατιστικής

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

Σύγχρονες μέθοδοι ανάλυσης δεδομένων επιβίωσης υψηλής διάστασης: Επιλογή μεταβλητών και μοντέλα πρόβλεψης

Νικόλαος Σγουρός

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην Εφαρμοσμένη Στατιστική

Πειραιάς
Σεπτέμβριος 2026

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣΕΣ του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. ..... συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Τα μέλη της Επιτροπής ήταν:

- Επίκουρος Καθηγητής Εμμανουήλ Ανδρουλάκης (Επιβλέπων)

- Καθηγητής Αλέξανδρος Καραγρηγορίου

- Επίκουρος Καθηγητής Φώτιος Σιάννης

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

UNIVERSITY OF PIRAEUS

School of Finance and Statistics

DEPARTMENT OF STATISTICS AND INSURANCE SCIENCE

POSTGRADUATE PROGRAM IN APPLIED STATISTICS

Modern methods for high-dimensional survival data analysis: Variable selection and predictive models

By

Nikolaos Sgouros

MSc Dissertation

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus, Greece

September 2026

Στην οικογένειά μου

και στους φίλους μου

Ευχαριστίες

Η παρούσα διπλωματική εργασία αποτελεί το κλείσιμο μιας όμορφης και ξεχωριστής περιόδου της ζωής μου, μέσα στην οποία είχα την ευκαιρία να γνωρίσω καλύτερα την επιστήμη που με ενδιαφέρει, να μάθω νέα πράγματα και να εξελιχθώ τόσο σε προσωπικό όσο και σε επιστημονικό επίπεδο. Κατά τη διάρκεια αυτού του ταξιδιού γνώρισα αξιόλογους ανθρώπους και συμφοιτητές, με τους οποίους στην πορεία δημιουργήθηκαν όμορφες φιλίες και μοιραστήκαμε πολλές στιγμές. Με την ολοκλήρωση αυτής της διπλωματικής εργασίας ολοκληρώνεται ένας σημαντικός κύκλος σπουδών και ξεκινά ένα νέο κεφάλαιο, γεμάτο νέες προκλήσεις, περισσότερη έρευνα και την ελπίδα ότι θα μπορέσω να συνεισφέρω στο χώρο της Ανάλυσης Επιβίωσης.

Θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα της διπλωματικής μου εργασίας και Επίκουρο Καθηγητή κύριο Εμμανουήλ Ανδρουλάκη, ο οποίος υπήρξε πολύτιμος καθοδηγητής σε όλη αυτή την πορεία. Τον ευχαριστώ για την εμπιστοσύνη που μου έδειξε, τις συμβουλές του, τις γνώσεις και τις εμπειρίες που μοιράστηκε μαζί μου, αλλά και για το χρόνο και την υποστήριξη που μου προσέφερε καθ' όλη τη διάρκεια της εκπόνησης της εργασίας. Θα ήθελα επίσης να ευχαριστήσω τον Καθηγητή κύριο Αλέξανδρο Καραγρηγορίου και τον Επίκουρο Καθηγητή κύριο Φώτιο Σιάννη, μέλη της τριμελούς επιτροπής, για το χρόνο που αφιέρωσαν στην ανάγνωση της παρούσας διπλωματικής εργασίας.

Τέλος, ένα μεγάλο ευχαριστώ θέλω να απευθύνω στην οικογένειά μου, η οποία βρίσκεται πάντα δίπλα μου, με στηρίζει και με ενθαρρύνει σε κάθε βήμα της πορείας μου. Η υπομονή, η αγάπη και η στήριξή τους αποτελέσαν για μένα σημαντική δύναμη όλα αυτά τα χρόνια. Θα ήθελα επίσης να ευχαριστήσω τους φίλους που γνώρισα κατά τη διάρκεια των σπουδών μου, αλλά και τους παιδικούς μου φίλους, οι οποίοι ήταν πάντα παρόντες, προσφέροντάς μου στήριξη, χαμόγελο και δύναμη στις δύσκολες στιγμές. Σας ευχαριστώ όλους για τη βοήθεια, την υποστήριξη και τις όμορφες στιγμές που μοιραστήκαμε.

Περίληψη

Η ανάλυση δεδομένων επιβίωσης υψηλής διάστασης αποτελεί ιδιαίτερα απαιτητικό στατιστικό πρόβλημα, καθώς ο αριθμός των διαθέσιμων επεξηγηματικών μεταβλητών μπορεί να είναι πολύ μεγαλύτερος από το μέγεθος του δείγματος, ενώ παράλληλα πρέπει να λαμβάνεται υπόψη η παρουσία λογοκριμένων παρατηρήσεων. Σε τέτοια περιβάλλοντα, η επιλογή των μεταβλητών με ουσιαστική προγνωστική πληροφορία και η ανάπτυξη μοντέλων με όσο το δυνατόν καλύτερη ικανότητα γενίκευσης αποτελούν βασικές προκλήσεις. Στην παρούσα διπλωματική εργασία παρουσιάζονται αρχικά οι βασικές έννοιες και τα κλασικά μοντέλα της Ανάλυσης Επιβίωσης και στη συνέχεια εξετάζονται σύγχρονες μέθοδοι για δεδομένα υψηλής διάστασης.

Σε περιβάλλοντα υψηλής διαστατικότητας ($p \gg n$), όπου η υπερπροσαρμογή και η πολυσυγγραμμικότητα περιορίζουν την προγνωστική ισχύ, προτείνεται μια στρατηγική δύο σταδίων, που περιλαμβάνει αρχικά μείωση της διάστασης μέσω feature screening (π.χ. RMST) και στη συνέχεια εφαρμογή ποινικοποιημένης παλινδρόμησης (regularization), όπως Broken Adaptive Ridge, Improved Adaptive Lasso, κ.α., για την ταυτόχρονη εκτίμηση και επιλογή μεταβλητών. Εξετάζονται επίσης σύγχρονες τεχνικές με εσωτερικούς μηχανισμούς ποινικοποίησης που εξασφαλίζουν αραιά (sparse) και ερμηνεύσιμα αποτελέσματα έχοντας υψηλής διάστασης δεδομένα επιβίωσης. Επιπλέον, εξετάζονται προηγμένες τεχνικές Μηχανικής Μάθησης για την ανίχνευση σύνθετων μη γραμμικών σχέσεων όπως Δέντρα Επιβίωσης (OST, OSST), Πλάγια Τυχαία Δάση Επιβίωσης (Oblique RSF), αλγόριθμοι Boosting, Survival SVM και προσεγγίσεις με σταθμισμένα βάρη Kaplan – Meier.

Η εργασία ολοκληρώνεται με την εφαρμογή επιλεγμένων τεχνικών σε ένα πραγματικό σύνολο δεδομένων επιβίωσης υψηλής διαστατικότητας, με στόχο τη δημιουργία βέλτιστου προγνωστικού μοντέλου βάσει του δείκτη συμφωνίας C – Index και την επιλογή των σημαντικότερων γονιδίων που συσχετίζονται με το χρόνο επιβίωσης.

Συνολικά, η εργασία αναδεικνύει τη σημασία της κατάλληλης επιλογής μεταβλητών, της ορθής διαδικασίας επικύρωσης και της συνδυαστικής αξιοποίησης στατιστικών και υπολογιστικών μεθόδων για την ανάπτυξη αξιόπιστων προγνωστικών μοντέλων σε δεδομένα επιβίωσης υψηλής διάστασης.

Abstract

The analysis of high – dimensional survival data constitutes a particularly challenging statistical problem, as the number of available explanatory variables can be far greater than the sample size, while the presence of censored observations must also be taken into account. In such settings, selecting variables with substantial prognostic information and developing models with the best possible generalization capability are key challenges. This thesis presents the fundamental concepts and classical models of Survival Analysis, and subsequently examines modern methods for high – dimensional data.

In high – dimensional environments ($p \gg n$), where overfitting and multicollinearity limit predictive performance, a two – stage strategy is proposed. This approach initially reduces dimensionality through feature screening (e.g., RMST), followed by the application of penalized regression (regularization) such as Broken Adaptive Ridge, Improved Adaptive Lasso, for simultaneous parameter estimation and variable selection. Modern techniques incorporating internal regularization mechanisms that produce sparse and interpretable results in high – dimensional survival data are also examined. Furthermore, advanced Machine Learning techniques are considered for detecting complex nonlinear relationships, including Survival Trees (OST, OSST), Oblique Random Survival Forests (Oblique RSF), Boosting algorithms, Survival SVM and approaches based on Kaplan – Meier weighted schemes.

The thesis concludes with the application of selected techniques to a real high – dimensional survival dataset, aiming to develop an optimal predictive model based on the Concordance Index (C-index) and identifies the most important genes associated with survival time.

Overall, this work highlights the importance of appropriate variable selection, proper validation procedures, and the combined utilization of statistical and computational methods to develop reliable prognostic models for high – dimensional survival data.

Πίνακας περιεχομένων

Κατάλογος Πινάκων

1ο Κεφάλαιο

3ο Κεφάλαιο

4ο Κεφάλαιο

Κατάλογος Σχημάτων

2ο Κεφάλαιο

3ο Κεφάλαιο

4ο Κεφάλαιο

Κατάλογος Συντομογραφιών

AFT	Accelerated Failure Time
AI	Artificial Intelligence
BAR	Broken Adaptive Ridge
BIC	Bayesian Information Criterion
c.d.f.	cumulative distribution function
c.h.f.	cumulative hazard function
CART	Classification and Regression Trees
CC-SVM	Constraint Classification - Support Vector Machine
CI Forests	Conditional Inference Forests
CI Trees	Conditional Inference Trees
C-Index	Concordance Index
CoxCS	Cox Conditional Screening
CURE	Censored Unbiased Regression Ensembles
CURT	Censored Unbiased Regression Trees
CUS	Cox Univariate Shrinkage
CUT	Censoring Unbiased Transformations
CV	cross – validation
ESF	empirical survival function
ELMCoxBAR	Extreme Learning Machine Cox Broken Adaptive Ridge
GAM	Generalized Additive Model
GEO	Gene Expression Omnibus
GBM	Gradient Boosting Method
GBMCI	Gradient Boosting Method Concordance Index
GLM	Generalized Linear Model
GP – PL	Gaussian Process Preference Learning
h.f.	hazard function
ICON	Information Criterion for Optimal Node splitting
IPCW	Inverse Probability of Censoring Weights
IPOD	Integrated Power Density
ISIS	Iterative Sure Independent Screening
IVM	Import Vector Machine
KM	Kaplan – Meier
LCIVs	Linear combinations of input variables
LGB	Light Gradient Boosting
LRS	Likelihood Ratio Statistic
LRT	Likelihood Ratio Test
LS-SVM	Least Squares Support Vector Machine
MAD	Median Absolute Deviation
MCP	Minimax Concave Penalty
MIO	Mixed Integer Optimization
ML	machine learning
MM	Majorization-Minimization

NCC	nested case – control
NP	Non-deterministic Polynomial
NPI	Nottingham Prognostic Index
OOB	Out – of – bag
ORSF	Oblique Random Survival Forests
OSCAR	Octagonal Shrinkage and Clustering Algorithm for Regression
OSST	Optimal Sparse Survival Tree
OST	Optimal Survival Trees
p.d.f.	probability density function
p.f.	probability function
PCA	Principal Component Analysis
PH	Proportional Hazards
PSO	Particle Swarm Optimization
RF	Random Forest
RIST	Recursive Partitioning Survival Tree
RMST	restricted mean survival time
RMSTSIS	Restricted Mean Survival Time Sure Independent Screening
RotSF	Rotation Survival Forest
RR	Restricted Ridge
RRotSF	Random Rotation Survival Forest
RSF	Random Survival Forest
s.f.	survival function
SCAD	Smoothly Clipped Absolute Deviation
SHAP	SHapley Additive exPlanation
SIS	Sure Independent Screening
SJS	Sure Joint Screening
SSVM	Survival Support Vector Machine
STS	Score Test Screening
SVC	Support Vector Classification
SVCR	Support Vector Censored Regression
SVM	Support Vector Machines
SVR	Support Vector Regression
UR	Unrestricted Ridge
VIMP	Variable Importance
WLSSVM	Weighted Least Squares Support Vector Machine
XGBLC	eXtreme Gradient Boosting and Lasso Cox algorithm
XGBoost	eXtreme Gradient Boosting
βλ.	βλέπε
ΕΛΠ	Έλεγχος Λόγου Πιθανοφανειών
κ.α.	και άλλα
κ.λπ.	και τα λοιπά
π.χ.	παραδείγματος χάριν

1ο Κεφάλαιο
Θεωρητικό Υπόβαθρο στην Ανάλυση Επιβίωσης

1.1 Εισαγωγή

Η Ανάλυση Επιβίωσης (Survival Analysis) είναι ένα σύνολο στατιστικών μεθόδων που μελετούν το χρόνο μέχρι την εμφάνιση ενός συγκεκριμένου συμβάντος. Το υπό μελέτη γεγονός μπορεί να αφορά την αστοχία ενός μηχανικού εξαρτήματος, το θάνατο ενός ασθενούς, την πτώχευση μιας εταιρείας ή οποιοδήποτε άλλο συμβάν που συμβαίνει σε μεταγενέστερο χρονικό σημείο. Η προέλευση της Ανάλυσης Επιβίωσης εντοπίζεται στον 17ο αιώνα, με τον John Graunt (1662), ο οποίος θεωρείται θεμελιωτής της δημογραφίας. Ήταν από τους πρώτους που κατασκεύασαν πίνακες επιβίωσης (life tables), με σκοπό τη μελέτη της θνησιμότητας. Λίγο αργότερα, ο Edmond Halley (1693) τους αξιοποίησε για την ανάπτυξη αναλογιστικών μοντέλων σε ασφαλιστικά πλαίσια. Από τα μέσα του 20ου αιώνα και εξής, η Ανάλυση Επιβίωσης εξελίχθηκε ραγδαία, με την εισαγωγή σημαντικών μεθόδων όπως οι Kaplan – Meier και Nelson – Aalen εκτιμητές, το Cox Proportional Hazards (Cox–PH) μοντέλο και τα Accelerated Failure Time (AFT) Models που αποτελούν μέχρι στιγμής το βασικό κορμό του πεδίου. Η ευρύτερη εφαρμογή της παρατηρείται κυρίως στο χώρο των βιοϊατρικών επιστημών και ειδικότερα στην ιατρική έρευνα, όπως στις κλινικές δοκιμές και στις επιδημιολογικές μελέτες. Οι γιατροί, σε συνεργασία με βιοστατιστικούς, μελετούν τη χρονική διάρκεια μέχρι την εμφάνιση σημαντικών συμβάντων, όπως ο θάνατος, η υποχώρηση της ασθένειας ή η ανταπόκριση σε κάποια θεραπεία.

Παρά την κυρίαρχη εφαρμογή στον ιατρικό τομέα, η Ανάλυση Επιβίωσης βρίσκει σημαντικές χρήσεις και σε άλλους κλάδους σύμφωνα με τον Collett (2023), όπως:

- Στα χρηματοοικονομικά, όπου εφαρμόζεται στη μελέτη του χρόνου μέχρι την πτώχευση επιχειρήσεων ή της διάρκειας της εγκριτικής διαδικασίας για τραπεζικά δάνεια.
- Στη μηχανική, όπου χρησιμοποιείται για την εκτίμηση της διάρκειας ζωής εξαρτημάτων ή προϊόντων, με στόχο είτε την παροχή εγγύησης στους πελάτες είτε την ενίσχυση της εμπορικής ανταγωνιστικότητας.
- Στη γεωργία, όπου μπορεί να αξιοποιηθεί για την πρόβλεψη του χρόνου εμφάνισης ασθενειών στα φυτά ή για την παρακολούθηση της ανάπτυξης καλλιεργειών.
- Στις κοινωνικές επιστήμες, όπου εφαρμόζεται για τη μελέτη ποικίλων κοινωνικών φαινομένων, όπως ο χρόνος μέχρι τη διάλυση ενός γάμου, ο χρόνος εύρεσης εργασίας μετά την αποφοίτηση, ή η χρονική διάρκεια παραμονής των νέων στο γονικό σπίτι.

Σ’ όλες αυτές τις περιπτώσεις, η Ανάλυση Επιβίωσης επιτρέπει όχι μόνο την εκτίμηση του χρόνου μέχρι την επέλευση ενός γεγονότος, αλλά και τη μελέτη παραγόντων που ενδεχομένως επηρεάζουν τη χρονική του κατανομή. Η παρούσα ενότητα εστιάζει κυρίως στη θεωρητική θεμελίωση της Ανάλυσης Επιβίωσης και στην εφαρμογή της σε ιατρικά δεδομένα, με στόχο τη μελέτη του χρόνου επιβίωσης και των παραγόντων που την επηρεάζουν.

1.2 Μελέτη κατανομής χρόνου Τ

Όπως αναφέρθηκε στην Ενότητα 1.1, ένας από τους κύριους στόχους της Ανάλυσης Επιβίωσης είναι η μελέτη του χρόνου μέχρι την εμφάνιση ενός καθορισμένου συμβάντος. Η τυχαία μεταβλητή που αναπαριστά το χρόνο αυτό, συμβολίζεται με το κεφαλαίο γράμμα T. Η μεταβλητή T μπορεί να είναι συνεχής, διακριτή ή και μικτή και λαμβάνει μη αρνητικές τιμές. Η συγκεκριμένη εργασία επικεντρώνεται στις δύο βασικές περιπτώσεις: τη συνεχή και τη διακριτή κατανομή του χρόνου στο διάστημα $[0, \infty)$. Στη συνέχεια, παρουσιάζονται οι βασικές συναρτήσεις που χρησιμοποιούνται στην Ανάλυση Επιβίωσης και στη στατιστική ερμηνεία των σχετικών δεδομένων.

1.2.1 Συνεχής Περίπτωση

Έστω ότι η τυχαία μεταβλητή T είναι συνεχής. Ορίζεται η συνάρτηση πυκνότητας πιθανότητας (probability density function, p.d.f.) ως $f(t)$, με πεδίο ορισμού $[0, \infty)$, για την οποία ισχύει:

$$\int_0^\infty f(t)dt = 1, \quad f(t) \geq 0 \ \ για\ κάθε\ t \geq 0, \ \ P(T = t) = 0$$

Η αθροιστική συνάρτηση κατανομής (cumulative distribution function, c.d.f.) ορίζεται ως:

$$F(t) = P(T \leq t) = \int_0^t f(x)dx \tag{1.1}$$

Ένας ισοδύναμος τύπος, που συνδέει την p.d.f. με τη c.d.f., είναι

$$f(t) = \frac{d}{dt}F(t) \tag{1.2}$$

Αντίστοιχα, η συνάρτηση επιβίωσης (survival function, s.f.) δίνει την πιθανότητα η μεταβλητή να ξεπεράσει τη χρονική στιγμή t:

$$S(t) = P(T > t) = \int_t^\infty f(x)dx = 1 - F(t) \tag{1.3}$$

Για τις συναρτήσεις (1.1), (1.2) ισχύει:

$$S(0) = 1 \Leftrightarrow F(0) = 0 \quad και \ \ S(\infty) = \lim_{t\to\infty} S(t) = 0 \Leftrightarrow F(\infty) = \lim_{t\to\infty} F(t) = 1$$

Η συνάρτηση πυκνότητας μπορεί να εκφραστεί και συναρτήσει της s.f. ως εξής:

$$f(t) = -\frac{d}{dt}S(t)$$

Το ποσοστημόριο τάξης $p \in (0,1)$ δηλαδή το $100p\%$ - ποσοστημόριο t_p ικανοποιεί τη σχέση:

$$F(t_p) = p \ \Leftrightarrow t_p = F^{-1}(p)$$

Μια σημαντική συνάρτηση στην Ανάλυση Επιβίωσης αποτελεί η συνάρτηση κινδύνου (hazard function, h.f.), η οποία εκφράζει το ρυθμό συμβάντος στο χρονικό διάστημα

$[t, t + \Delta t)$, μεταξύ αυτών που βρίσκονται σε κίνδυνο στην αρχή του διαστήματος, δεν είναι πιθανότητα και ορίζεται ως:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t \mid T \ge t)}{\Delta t} = \frac{f(t)}{S(t)} \tag{1.4}$$

Ισοδύναμα, εκφράζεται και ως:

$$h(t) = -\frac{d}{dt} \log S(t) \tag{1.5}$$

Από τον παραπάνω ορισμό, η s.f. διατυπώνεται και ως εξής:

$$S(t) = \exp\left(-\int_0^t h(x)dx\right)$$

και επομένως, η p.d.f. γράφεται ως:

$$f(t) = h(t) \exp\left(-\int_0^t h(x)dx\right)$$

Τέλος, ορίζεται η αθροιστική συνάρτηση κινδύνου (cumulative hazard function, c.h.f.) ως:

$$H(t) = \int_0^t h(x)d(x)$$

η οποία συνδέεται με την s.f. ως ακολούθως:

$$S(t) = \exp(-H(t)) \tag{1.6}$$

1.2.2 Διακριτή Περίπτωση

Υπάρχουν περιπτώσεις όπου η τυχαία μεταβλητή T μπορεί να είναι διακριτή και να λαμβάνει τιμές $0 \le t_1 < t_2 < \cdots$. Ορίζεται η συνάρτηση πιθανότητας (probability function, p.f.) της T ως $f(t)$ με πεδίο ορισμού $B = \{t_1, t_2, \dots\} \subseteq [0, \infty)$, για την οποία ισχύει:

$$\sum_{t \in B} f(t) = 1, \quad f(t) = P(T = t) > 0 \ \ \forall t \in B, \quad f(t) = 0, \forall t \notin B \tag{1.7}$$

Η c.d.f. ορίζεται αντιστοίχως, όπως και στη συνεχή περίπτωση, ως εξής:

$$F(t) = P(T < t) = \sum_{\substack{x \in B \\ x < t}} f(x) \tag{1.8}$$

Η s.f. μιας διακριτής κατανομής ορίζεται ως:

$$S(t) = P(T \ge t) = \sum_{\substack{x \in B \\ x \ge t}} f(x) = 1 - F(t-) \tag{1.9}$$

Από τον τύπο (1.9) διαπιστώνεται ότι η συνάρτηση επιβίωσης είναι μία αριστερά – συνεχής κλιμακωτή συνάρτηση η οποία αλλάζει βήμα σε κάθε $t \in B$. Επιπρόσθετα, για τις συναρτήσεις (1.8), (1.9) ισχύουν οι ιδιότητες των c.d.f., s.f. που αναφέρονται στην ενότητα 1.2.1. Η h.f. διακριτής τυχαίας μεταβλητής T ορίζεται ως:

$$h(t) = P(T = t \mid T \ge t) = \frac{f(t)}{S(t)}, \forall t \in B$$

και επειδή ισχύει ότι $f(t) = S(t) - S(t+1)$, ο τύπος (1.5) μπορεί να γραφεί ως:

$$h(t) = 1 - \frac{S(t+1)}{S(t)}, \forall t \in B$$

Επομένως, ένας ισοδύναμος τύπος του (1.9) για τη διακριτή περίπτωση είναι

$$S(t) = \prod_{\substack{x<t \\ x\in B}} \big(1 - h(x)\big)$$

Η c.h.f. μπορεί να εκφραστεί με δύο διαφορετικούς τρόπους. Ο πρώτος που ισχύει τόσο στη συνεχή όσο και στη διακριτή περίπτωση είναι:

$$H(t) = -\log S(t) \tag{1.10}$$

Ο δεύτερος τρόπος ορισμού, ο οποίος χρησιμοποιείται κυρίως στη διακριτή περίπτωση και αποτελεί προσεγγιστική μορφή της (1.10), δίνεται από:

$$H(t) = \sum_{\substack{x<t \\ x\in B}} h(x)$$

1.2.3 Πιθανοθεωρητικές κατανομές για την ανάλυση του χρόνου T

Στην Ανάλυση Επιβίωσης, η μοντελοποίηση του χρόνου μέχρι την επέλευση ενός γεγονότος T (όπως θάνατος, αστοχία, αποχώρηση, κ.α.) αποτελεί θεμελιώδες ζήτημα. Σε ορισμένες περιπτώσεις, η μορφή της κατανομής T με κατάλληλους ελέγχους καλής προσαρμογής και ορθή στατιστική ανάλυση είναι εύκολο να εκτιμηθεί. Ωστόσο, συχνά δεν είναι δυνατός ο σαφής προσδιορισμός της υποκείμενης κατανομής, οπότε επιλέγονται μη παραμετρικές ή ημιπαραμετρικές μέθοδοι οι οποίες αναλύονται σε επόμενες υποενότητες του Κεφαλαίου. Σε κάθε περίπτωση όταν είναι εφικτή η παραμετρική προσέγγιση, η χρήση κατάλληλης θεωρητικής κατανομής μπορεί να οδηγήσει σε ακριβέστερα και πιο ερμηνεύσιμα μοντέλα. Βάσει της διεθνούς βιβλιογραφίας και της εμπειρικής εφαρμογής, οι ακόλουθες κατανομές έχουν αποδειχθεί ιδιαίτερα χρήσιμες στη μοντελοποίηση χρόνων επιβίωσης.

Εκθετική κατανομή – $\boldsymbol{Exp(\theta = 1/\lambda)}$

Η εκθετική κατανομή με μέση τιμή $\theta > 0$ και παράμετρο $\lambda > 0$ είναι η απλούστερη κατανομή που χρησιμοποιείται στον κλάδο. Χαρακτηρίζεται από σταθερή $h(t) = \lambda$, και εφαρμόζεται κυρίως σε περιπτώσεις όπου ο ρυθμός αποτυχίας δεν εξαρτάται από το χρόνο t.

Κατανομή Weibull - $\boldsymbol{Weib(\lambda, \gamma)}$

Η Weibull κατανομή αποτελεί μια γενίκευση της Εκθετικής κατανομής με παραμέτρους κλίμακας λ και σχήματος $\gamma > 0$. Για $\gamma = 1$, η κατανομή εκφυλίζεται στην εκθετική. Η $h.f.$ είναι μονότονη: αυξανόμενη για $\gamma > 1$ και φθίνουσα για $0 < \gamma < 1$. Η ευελιξία αυτή την καθιστά από τις πλέον δημοφιλείς επιλογές στην Παραμετρική Ανάλυση Επιβίωσης.

Κατανομή Ακραίων Τιμών - $EV(u, \beta)$

Η κατανομή Ακραίων Τιμών, γνωστή και ως κατανομή Gumbel (τύπου Ι), ορίζεται από τις παραμέτρους θέσης $u \in \mathbb{R}$ και κλίμακας $\beta > 0$. Ένα σημαντικό αποτέλεσμα από τη θεωρία των Πιθανοτήτων είναι ότι αν $T \sim Weib(\lambda_{Weib}, \gamma_{Weib})$ τότε $\log T \sim EV\left(u = -\log\frac{\lambda_{Weib}}{\gamma}, \beta = \frac{1}{\gamma_{Weib}}\right)$. Η σχέση αυτή αξιοποιείται στα παραμετρικά μοντέλα τύπου AFT στα οποία μελετάται η κατανομή του $\log T$ αντί του ίδιου του T.

Λογαριθμοκανονική Κατανομή - $Log - Normal(\mu, \sigma^2)$

Η κατανομή αυτή με παραμέτρους $\mu \in \mathbb{R}$ και $\sigma^2 > 0$, είναι κατάλληλη όταν το $\log T \sim N(\mu, \sigma^2)$. Η $Log - Normal(\mu, \sigma^2)$ έχει εφαρμογές σε πολλές επιστήμες, μεταξύ των οποίων η μηχανική, η ιατρική και οι κοινωνικές επιστήμες. Η h.f. είναι μη μονοτονική: ξεκινά από το μηδέν για $t = 0$ αυξάνεται, φτάνει σε ένα σημείο μέγιστο και έπειτα μειώνεται καθώς $t \to \infty$. Η μορφή της εξαρτάται κυρίως από την τιμή της παραμέτρου σ^2, γεγονός που επιτρέπει την ευέλικτη μοντελοποίηση τόσο μικρών όσο και μεγάλων χρόνων επιβίωσης.

Λογαριθμολογιστική Κατανομή - $Log - Logist(\alpha, \beta)$)

Η Λογαριθμολογιστική κατανομή με παραμέτρους κλίμακας $\alpha > 0$ και σχήματος $\beta > 0$ προκύπτει όταν το $\log T \sim Logistic\left(u = \log a, b = \frac{1}{\beta}\right)$. Παρουσιάζει παρόμοια μορφή h.f. με τη Λογαριθμοκανονική. Ωστόσο, για τιμές $\beta < 1$, η h.f. είναι φθίνουσα ως προς το χρόνο t.

Κατανομή Γάμμα - $Ga(k, \frac{1}{\lambda})$

Η Gamma Κατανομή μπορεί να θεωρηθεί κι αυτή γενίκευση της Εκθετικής κατανομής (για $k = 1$) και ορίζεται από τις παραμέτρους σχήματος $k > 0$ και κλίμακας $1/\lambda$. Σύμφωνα με τη θεωρία, το άθροισμα k ανεξάρτητων μεταβλητών που ακολουθούν εκθετική κατανομή με παράμετρο λ ακολουθεί την κατανομή $Ga\left(k, \frac{1}{\lambda}\right)$. Η h.f. εμφανίζει διαφορετική συμπεριφορά ανάλογα με την τιμή της παραμέτρου σχήματος. Για $k < 1$ είναι φθίνουσα συνάρτηση του t με $\lim_{t\to 0} h(t) = \infty$ και $\lim_{t\to\infty} h(t) = \lambda$ και για $k > 1$ είναι αυξανόμενη συνάρτηση του t με $h(0) = 0$ και $\lim_{t\to\infty} h(t) = \lambda$. Οι παραπάνω κατανομές, καθεμία με τα δικά της χαρακτηριστικά και περιορισμούς, προσφέρουν ένα ευρύ φάσμα επιλογών για την παραμετρική προσέγγιση στην Ανάλυση Επιβίωσης. Η κατάλληλη επιλογή εξαρτάται τόσο από τα εμπειρικά δεδομένα όσο και από τη θεωρητική κατανόηση της μορφής της h.f. του υπό μελέτη φαινομένου. Τέλος, στον Πίνακα 1.1 παρουσιάζονται αναλυτικά κάποιες βασικές συναρτήσεις που αναφέρθηκαν στην ενότητα 1.2 για κάθε παραπάνω μεταβλητή αναφέροντας και ακόμη δύο κατανομές Logistic και Normal που θα χρησιμοποιηθούν στα μοντέλα επιταχυνόμενου χρόνου αστοχίας/αποτυχίας (Accelerated Failure Time – AFT models).

Πίνακας 1.1 ***Βασικές Συναρτήσεις των 8 κατανομών***

Κατανομή	**Συνάρτηση πυκνότητας πιθανότητας** $\boldsymbol{f(t)}$ $\boldsymbol{(p.d.f.)}$	**Αθροιστική συνάρτηση κατανομής** $\boldsymbol{F(t)}$ $\boldsymbol{(c.d.f.)}$	**Συνάρτηση Επιβίωσης** $\boldsymbol{S(t)}$ $\boldsymbol{(s.f.)}$	**Συνάρτηση κινδύνου** $\boldsymbol{h(t)}$ $\boldsymbol{(h.f.)}$	**Μέση Τιμή** $\boldsymbol{E(T)}$	**Διακύμανση** $\boldsymbol{Var(T)}$
$\boldsymbol{Exp(\theta = 1/\lambda)}$ $\theta, \lambda > 0$ $t \geq 0$	$\theta^{-1}e^{-\frac{t}{\theta}}$	$1 - e^{-\frac{t}{\theta}}$	$e^{-\frac{t}{\theta}}$	$\frac{1}{\theta} = \lambda$	$\frac{1}{\theta} = \lambda$	$\frac{1}{\theta^2} = \lambda^2$
$\boldsymbol{Weib(\lambda, \beta)}$ $\lambda, \beta > 0$ $t > 0$	$\lambda\gamma t^{\gamma-1}e^{-\lambda t^{\gamma}}$	$1 - e^{-\lambda t^{\gamma}}$	$e^{-\lambda t^{\gamma}}$	$\lambda\gamma t^{\gamma-1}$	$\frac{1}{\lambda^{1/\gamma}}\Gamma\left(1+\frac{1}{\gamma}\right)$	$\frac{1}{\lambda^{2/\gamma}}\left[\Gamma\left(1+\frac{2}{\gamma}\right) - \Gamma\left(1+\frac{1}{\gamma}\right)^2\right]$
$\boldsymbol{Ga\left(k, \frac{1}{\lambda}\right)}$ $\kappa, \lambda > 0$ $t > 0$	$\frac{\lambda^k t^{k-1}e^{-\lambda t}}{\Gamma(k)}$	$I(k,t)$	$1 - I(k,t)$	$\frac{\lambda^k t^{k-1}e^{-\lambda t}}{\Gamma(k)(1-I(k,t))}$	$\frac{\kappa}{\lambda}$	$\frac{\kappa}{\lambda^2}$
$\boldsymbol{EV(u, \beta)}$ $u \in \mathbb{R}$, $\beta > 0$ $t \in \mathbb{R}$	$\frac{e^{\left[\frac{t-u}{\beta} - e^{\left(\frac{t-u}{\beta}\right)}\right]}}{\beta}$	$1 - e^{-e^{\left(\frac{t-u}{\beta}\right)}}$	$e^{-e^{\left(\frac{t-u}{\beta}\right)}}$	$\frac{e^{\frac{t-u}{\beta}}}{\beta}$	$u + \gamma\beta$	$\left(\frac{\pi^2}{6}\right)\beta^2$
$\boldsymbol{Normal(\mu, \sigma^2)}$ $\mu \in \mathbb{R}, \sigma^2 > 0$ $t \in \mathbb{R}$	$\frac{1}{\sigma\sqrt{2\pi}}e^{\left[-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right]}$	$\Phi\left(\frac{t-\mu}{\sigma}\right)$	$1 - \Phi\left(\frac{t-\mu}{\sigma}\right)$	$\frac{\frac{1}{\sigma\sqrt{2\pi}}e^{\left[-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right]}}{1-\Phi\left(\frac{t-\mu}{\sigma}\right)}$	μ	σ^2
$\boldsymbol{Log - Normal(\mu, \sigma^2)}$ $\mu \in \mathbb{R}, \sigma^2 > 0$ $t > 0$	$\frac{1}{\sigma t\sqrt{2\pi}}e^{\left[-\frac{1}{2}\left(\frac{\log t-\mu}{\sigma}\right)^2\right]}$	$\Phi\left(\frac{\log t-\mu}{\sigma}\right)$	$1 - \Phi\left(\frac{\log t-\mu}{\sigma}\right)$	$\frac{\frac{1}{\sigma t\sqrt{2\pi}}e^{\left[-\frac{1}{2}\left(\frac{\log t-\mu}{\sigma}\right)^2\right]}}{1-\Phi\left(\frac{\log t-\mu}{\sigma}\right)}$	$e^{\left(\mu+\frac{\sigma^2}{2}\right)}$	$(e^{\sigma^2}-1)(e^{2\mu+\sigma^2})$

Κατανομή	**Συνάρτηση πυκνότητας πιθανότητας** $\boldsymbol{f(t)}$ $\boldsymbol{(p.d.f.)}$	**Αθροιστική συνάρτηση κατανομής** $\boldsymbol{F(t)}$ $\boldsymbol{(c.d.f.)}$	**Συνάρτηση Επιβίωσης** $\boldsymbol{S(t)}$ $\boldsymbol{(s.f.)}$	**Συνάρτηση κινδύνου** $\boldsymbol{h(t)}$ $\boldsymbol{(h.f.)}$	**Μέση Τιμή** $\boldsymbol{E(T)}$	**Διακύμανση** $\boldsymbol{Var(T)}$
$\boldsymbol{Logistic(u,b)}$ $u \in \mathbb{R}, s > 0$ $t \in \mathbb{R}$	$\frac{e^{-\frac{t-u}{b}}}{s\left(1+e^{-\frac{t-u}{b}}\right)^2}$	$\frac{1}{1+e^{-\frac{t-u}{b}}}$	$\frac{1}{1+e^{\frac{t-u}{b}}}$	$\frac{1}{s\left(1+e^{-\frac{t-u}{b}}\right)}$	u	$\frac{\pi^2 b^2}{3}$
$\boldsymbol{Log-Logistic(\alpha,\beta)}$ $\alpha, \beta > 0$ $t > 0$	$\frac{\left(\frac{\beta}{\alpha}\right)\left(\frac{t}{\alpha}\right)^{\beta-1}}{\left(1+\left(\frac{t}{\alpha}\right)^{\beta}\right)^2}$	$1-\left(1+\left(\frac{t}{\alpha}\right)^{\beta}\right)^{-1}$	$\left(1+\left(\frac{t}{\alpha}\right)^{\beta}\right)^{-1}$	$\frac{\left(\frac{\beta}{\alpha}\right)\left(\frac{t}{\alpha}\right)^{\beta-1}}{1+\left(\frac{t}{\alpha}\right)^{\beta}}$	$\alpha\Gamma\left(1+\frac{1}{\beta}\right)\Gamma\left(1-\frac{1}{\beta}\right)$ [Υπό την προϋπόθεση ότι $\beta > 1$]	$\alpha^2\left(\frac{\pi\beta}{\sin\left(\frac{\pi}{\beta}\right)}\right)^2\left(\frac{1}{\beta-2}-\left(\frac{1}{\beta-1}\right)^2\right)$ [Υπό την προϋπόθεση ότι $\beta > 2$]

Πίνακας 1.2 ***Βοηθητικές συναρτήσεις για τον Πίνακα 1.1***

Όπου γ=0.5772 (σταθερά Euler's)	*Όπου π=3.14159 (μαθηματική σταθερά)*	$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du$	$I(k,t) = \frac{1}{\Gamma(k)} \int_0^t u^{k-1} e^{-u} du$	$\Gamma(\kappa) = \int_0^{\infty} u^{k-1} e^{-u} du, k > 0$

Πίνακας 1.3 ***Μαθηματικές σχέσεις και μετατροπές μεταξύ των βασικών συναρτήσεων της Ανάλυσης Επιβίωσης***

Συνάρτηση	$\boldsymbol{f(t)}$ **(p.d.f.)**	$\boldsymbol{F(t)}$ **(c.d.f.)**	$\boldsymbol{S(t)}$ **(s.f.)**	$\boldsymbol{h(t)}$ **(h.f.)**	$\boldsymbol{H(t)}$ **(c.h.f.)**
$\boldsymbol{f(t)}$ **(p.d.f.)**	=	$F'(t)$	$-S'(t)$	$\exp\left\{-\int_0^t h(x)dx\right\}$	$H'(t)\exp\{-H(t)\}$
$\boldsymbol{F(t)}$ **(c.d.f.)**	$\int_0^t f(x)dx$	=	$1-S(t)$	$1-\exp\left\{-\int_0^t h(x)dx\right\}$	$1-\exp\{-H(t)\}$
$\boldsymbol{S(t)}$ **(s.f.)**	$\int_t^\infty f(x)dx$	$1-F(t)$	=	$\exp\left\{-\int_0^t h(x)dx\right\}$	$\exp\{-H(t)\}$
$\boldsymbol{h(t)}$ **(h.f.)**	$\frac{f(t)}{\int_t^\infty f(x)dx}$	$\frac{F'(t)}{1-F(t)}$	$-(\log S(t))'$	=	$H'(t)$
$\boldsymbol{H(t)}$ **(c.h.f.)**	$-\log\left(\int_t^\infty f(x)dx\right)$	$-\log\big(1-F(t)\big)$	$-\log S(t)$	$\int_0^t h(x)dx$	=

1.3 Λογοκρισία/Αποκοπή των Δεδομένων

Η Ανάλυση Επιβίωσης αποτελεί μία εξειδικευμένη μορφή στατιστικής ανάλυσης, η οποία διαφοροποιείται από τις κλασικές μεθόδους καθώς λαμβάνει υπόψη την πιθανότητα λογοκρισίας στα δεδομένα. Λογοκρισία (censoring) σε μια στατιστική μελέτη σημαίνει ότι ο αναλυτής δεν γνωρίζει την ακριβή χρονική στιγμή κατά την οποία συνέβη το γεγονός που τον ενδιαφέρει (π.χ. θάνατος, εμφάνιση ασθένειας, αποτυχία ενός μηχανισμού), αλλά διαθέτει μερική πληροφορία μέχρι ένα ορισμένο χρονικό σημείο, ή από ένα χρονικό σημείο κι ύστερα ή για ένα συγκεκριμένο χρονικό διάστημα. Υπάρχουν διάφοροι τύποι λογοκρισίας και ανάλογα τον τύπο διαμορφώνονται αναλόγως και οι συναρτήσεις πιθανοφάνειας που χρησιμοποιούνται για την ορθή στατιστική ανάλυση των εν λόγω δεδομένων. Η εκτίμηση των παραμέτρων του μοντέλου βασίζεται στη μεγιστοποίηση της αντίστοιχης συνάρτησης πιθανοφάνειας (π.χ. σε παραμετρικά μοντέλα όπως Exponential, Weibull, κ.λπ.), προκειμένου να εξαχθούν αξιόπιστες εκτιμήσεις για το χρόνο επιβίωσης του υπό μελέτη φαινομένου.

Οι τρεις βασικές κατηγορίες λογοκρισίας είναι η δεξιά λογοκρισία (right censoring), η αριστερή λογοκρισία (left censoring) και η λογοκρισία σε διάστημα (interval censoring). Οι δύο πρώτες μορφές λογοκρισίας (δεξιά και αριστερή) μπορούν να θεωρηθούν ειδικές περιπτώσεις της λογοκρισίας σε διάστημα όπου το χρονικό διάστημα στο οποίο είναι γνωστό ότι συνέβη το γεγονός εκφυλίζεται σε ημι-άπειρα όρια (π.χ. από ένα χρόνο κι ύστερα, ή μέχρι έναν συγκεκριμένο χρόνο). Πέραν αυτών των βασικών τύπων, υπάρχουν και άλλες δύο μορφές λογοκρισίας, όπως είναι η ανεξάρτητη τυχαία λογοκρισία και η εξαρτημένη από το χρόνο λογοκρισία, οι οποίες και θα αναλυθούν στη συνέχεια της υποενότητας. Επιπλέον, για την πληρέστερη κατανόηση των παραπάνω εννοιών, είναι σημαντική η διάκριση μεταξύ περιπτώσεων με και χωρίς δεσμούς/ισοπαλίες (ties). Η απουσία δεσμών σημαίνει ότι όλοι οι χρόνοι επιβίωσης στο δείγμα είναι μοναδικοί (δεν υπάρχουν παρατηρήσεις με τον ίδιο χρόνο). Αντιθέτως, η παρουσία δεσμών υποδηλώνει ότι υπάρχουν τουλάχιστον δύο ίσοι παρατηρούμενοι χρόνοι στο δείγμα, γεγονός που πρέπει να ληφθεί υπόψη κατά την ανάλυση.

1.3.1 Δεξιά Λογοκρισία

Σύμφωνα με το Lawless (2003), στον οποίο βασίζονται οι ορισμοί που ακολουθούν, η δεξιά λογοκρισία αποτελεί μια από τις πλέον συχνά εμφανιζόμενες περιπτώσεις λογοκριμένων παρατηρήσεων σε μελέτες με πραγματικά δεδομένα. Στην περίπτωση δεξιάς λογοκρισίας, για κάθε άτομο είναι γνωστό μόνο ένα κατώτερο όριο του χρόνου επιβίωσης, δηλαδή είναι γνωστό ότι το γεγονός ενδιαφέροντος δεν έχει πραγματοποιηθεί μέχρι και ένα συγκεκριμένο χρονικό σημείο, χωρίς να είναι γνωστό εάν και πότε πραγματοποιήθηκε στη συνέχεια. Με άλλα λόγια, η πραγματική διάρκεια του χρόνου μέχρι το γεγονός υπερβαίνει το χρονικό σημείο λογοκρισίας αλλά η ακριβής τιμή παραμένει άγνωστη. Αιτίες που μπορεί να οδηγήσουν σε δεξιά λογοκρισία περιλαμβάνουν πρόωρο τερματισμό της μελέτης, πριν προλάβει να

παρατηρηθεί το γεγονός σε όλα τα άτομα, μη προγραμματισμένη αποχώρηση συμμετεχόντων λόγω αλλαγής τοποθεσίας ή κατοικίας, λόγω προσωπικών λόγων (π.χ. αμφιβολίες, φόβος, απώλεια ενδιαφέροντος) ή άλλων εξωτερικών παραγόντων. Η σωστή αναγνώριση και χειρισμός της δεξιάς λογοκρισίας είναι κρίσιμη για την αξιόπιστη εκτίμηση των παραμέτρων σε μοντέλα επιβίωσης, καθώς και για την ακρίβεια των συμπερασμάτων που προκύπτουν από την ανάλυση.

1.3.1.1 Δεξιά Λογοκρισία Τύπου I

Η δεξιά λογοκρισία τύπου I (right censoring type I) εμφανίζεται όταν κάθε παρατήρηση – άτομο έχει έναν προκαθορισμένο χρόνο λογοκρισίας, που συμβολίζεται ως C_i. Αν ο πραγματικός χρόνος εμφάνισης του γεγονότος T_i είναι μικρότερος από το χρόνο λογοκρισίας C_i ($T_i < C_i$) τότε παρατηρείται ο πραγματικός χρόνος T_i και θεωρείται ότι το $i-$ άτομο έχει πλήρη χρόνο. Αν όμως $T_i \geq C_i$ τότε το γεγονός δεν έχει συμβεί μέχρι το χρόνο λογοκρισίας και η παρατήρηση κατατάσσεται ως λογοκριμένη. Στην περίπτωση αυτή καταγράφεται ως $T_i = C_i +$ δηλαδή ο χρόνος C_i με ένδειξη λογοκρισίας (το σύμβολο «+» δηλώνει ότι ο πραγματικός χρόνος υπερβαίνει το C_i).

Για παράδειγμα, δεξιά λογοκρισία τύπου I παρατηρείται όταν σε μία μελέτη $n = 100$ ατόμων χορηγείται ένα φάρμακο για την καταπολέμηση μιας ασθένειας και οι ασθενείς παρακολουθούνται για ένα προκαθορισμένο χρονικό διάστημα, έστω δύο χρόνια, με σκοπό να μελετηθεί ο χρόνος υποτροπής της ασθένειας. Κατά τη διάρκεια της μελέτης, σε ορισμένα άτομα ενδέχεται να παρατηρηθεί υποτροπή και σ' αυτές τις περιπτώσεις καταγράφεται ο ακριβής χρόνος εμφάνισης του γεγονότος. Αντίθετα, για τα άτομα που δεν εμφανίζουν υποτροπή εντός των 2 χρόνων που διαρκεί η μελέτη, απλώς είναι γνωστό ότι δεν υπήρξε υποτροπή μέχρι το τέλος της παρακολούθησης. Οι παρατηρήσεις αυτές θεωρούνται δεξιά λογοκριμένες. Η συνάρτηση πιθανοφάνειας που χρησιμοποιείται για την ανάλυση δεδομένων με δεξιά λογοκρισία τύπου I έχει την ακόλουθη μορφή:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(t_i; \boldsymbol{\theta})^{\delta_i} S(t_i+; \boldsymbol{\theta})^{1-\delta_i} \qquad (1.11)$$

όπου

- $\boldsymbol{\theta}$: είναι το διάνυσμα των παραμέτρων της κατανομής του χρόνου T**,**
- $t_i = \min(T_i, C_i)$ είναι ο παρατηρούμενος χρόνος για το $i-$ άτομο
- $\delta_i = I(T_i \leq C_i)$ είναι μία δείκτρια συνάρτηση που λαμβάνει τιμή 1 αν ο πραγματικός χρόνος συμβάντος T_i είναι μικρότερος ή ίσος του χρόνου λογοκρισίας C_i (δηλαδή όταν το γεγονός έχει παρατηρηθεί) και 0 σε αντίθετη περίπτωση (όταν η παρατήρηση είναι λογοκριμένη).

Στη συνάρτηση (1.11), για τις μη λογοκριμένες παρατηρήσεις ($\delta_i = 1$) συμμετέχει η $f(t)$ καθώς το γεγονός έχει συμβεί και ο χρόνος του είναι γνωστός, ενώ για τις λογοκριμένες παρατηρήσεις ($\delta_i = 0$) συμμετέχει η συνάρτηση επιβίωσης $S(t+) = P(T > C)$ δηλαδή η πιθανότητα το γεγονός να μην έχει συμβεί έως το χρόνο C. Στην περίπτωση που ο χρόνος T είναι συνεχής, ισχύει ότι $S(t+) = S(t)$. Η συγκεκριμένη μορφή της πιθανοφάνειας, όπως και στις επόμενες περιπτώσεις, είναι

θεμελιώδους σημασίας για την εκτίμηση των παραμέτρων $\boldsymbol{\theta}$ σε μοντέλα επιβίωσης, καθώς επιτρέπει τη συνδυαστική αξιοποίηση τόσο των παρατηρούμενων όσο και των λογοκριμένων δεδομένων, χωρίς να αγνοεί καμία παρατήρηση.

1.3.1.2 Δεξιά Λογοκρισία Τύπου ΙΙ

Στη δεξιά λογοκρισία τύπου ΙΙ, υπάρχει ένα τυχαίο δείγμα n ατόμων και παρατηρούνται μόνο οι r πρώτοι χρόνοι ζωής, έστω $t_{(1)} < t_{(2)} < \cdots < t_{(r)}$. Αυτό το είδος λογοκρισίας δεν εφαρμόζεται τόσο συχνά στην πράξη αλλά περισσότερο σε θεωρητικά προβλήματα, ωστόσο αν εφαρμοσθούν στην πραγματικότητα γίνεται λόγω υψηλού κόστους χρόνου/παρακολούθησης των ατόμων της μελέτης. Ο αριθμός r είναι ένας προκαθορισμένος ακέραιος μεταξύ του 1 και n, ο οποίος επιλέγεται πριν την έναρξη της μελέτης. Αυτός ο τύπος λογοκρισίας έχει το χαρακτηριστικό ότι η μελέτη τερματίζεται όταν παρατηρηθεί ο χρόνος ζωής του r −οστού ατόμου, και συνεπώς, ο χρόνος λήξης της μελέτης δεν είναι γνωστός εκ των προτέρων. Με την προϋπόθεση συνεχούς κατανομής του χρόνου ζωής T, αγνοείται και η πιθανότητα δεσμών. Έχοντας παρατηρήσει τους πρώτους r χρόνους ζωής $t_{(1)} < t_{(2)} < \cdots < t_{(r)}$, η συνάρτηση πιθανοφάνειας (likelihood function) για το δείγμα δίνεται από τον τύπο:

$$L(\boldsymbol{\theta}) = \frac{n!}{(n-r)!}\left\{\prod_{i=1}^{r} f(t_{(i)};\boldsymbol{\theta})\right\} S\left(t_{(r)};\boldsymbol{\theta}\right)^{n-r}$$

Αυτού του είδους λογοκρισία μπορεί να έχει εφαρμογή στη Μηχανική. Έστω ότι ένας μηχανικός αξιοπιστίας θέλει να μελετήσει τη διάρκεια ζωής 10 ιδίων ηλεκτρονικών εξαρτημάτων ($n = 10$) που τίθενται σε λειτουργία ταυτόχρονα. Αντί να παρακολουθεί όλα τα εξαρτήματα μέχρι να αποτύχουν, αποφασίζει εκ των προτέρων να τερματίσει τη δοκιμή όταν αποτύχουν τα 5 πρώτα εξαρτήματα ($r = 5$). Καθώς προχωρά η δοκιμή, καταγράφει τους ακριβείς χρόνους αποτυχίας των πρώτων 5 εξαρτημάτων, ενώ για τα υπόλοιπα 5 γνωρίζει ότι δεν έχουν αποτύχει μέχρι το χρόνο που συνέβη η $5^{η}$ αποτυχία. Οι χρόνοι ζωής των τελευταίων 5 εξαρτημάτων θεωρούνται δεξιά λογοκριμένοι, καθώς δεν είναι γνωστός ο χρόνος αποτυχίας τους (που ίσως προκύψει στο μέλλον) – παρά μόνο ότι δεν απέτυχαν μέχρι το χρόνο της $5^{ης}$ αποτυχίας.

Τέλος, στη συγκεκριμένη κατηγορία λογοκρισίας υπάρχει και μια γενικευμένη μορφή, γνωστή ως προοδευτική δεξιά λογοκρισία τύπου ΙΙ (progressive Type II right censoring). Στο σχεδιασμό αυτό, η μελέτη ξεκινά με n άτομα, αλλά η απομάκρυνση συμμετεχόντων γίνεται σταδιακά και με προγραμματισμένο τρόπο σε κάθε στάδιο της παρακολούθησης. Συγκεκριμένα, στην αρχή της μελέτης παρατηρούνται οι πρώτοι r_1 χρόνοι ζωής, και από τα $n - r_1$ άτομα που απομένουν, αποσύρονται n_1 άτομα, αφήνοντας $n - r_1 - n_1$ να συνεχίσουν. Σε επόμενο στάδιο, παρατηρούνται οι επόμενοι r_2 χρόνοι ζωής αυτών των ατόμων, και μετά αποσύρονται επιπλέον n_2 άτομα. Η διαδικασία αυτή συνεχίζεται για προκαθορισμένο αριθμό σταδίων μέχρις ότου ολοκληρωθεί η μελέτη. Η προοδευτική λογοκρισία προσφέρει μεγαλύτερη ευελιξία στο σχεδιασμό, ιδιαίτερα σε πειράματα που έχουν μεγάλο κόστος ή διάρκεια. Ωστόσο, η αυξημένη πολυπλοκότητα στη διαδικασία καθιστά πιο δύσκολη

τόσο τη διατύπωση της συνάρτησης πιθανοφάνειας, όσο και την εφαρμογή της σε πραγματικά προβλήματα, ιδιαίτερα όταν απαιτείται παραμετρική εκτίμηση.

1.3.2 Αριστερή Λογοκρισία

Σύμφωνα με τον Collett (2023), η αριστερή λογοκρισία (left censoring) αποτελεί μια μορφή λογοκρισίας που συναντάται σπανιότερα σε εφαρμοσμένες μελέτες (στην πράξη), ενώ η χρήση της περιορίζεται κυρίως σε θεωρητικό επίπεδο, ώστε να διαμορφωθεί ένα πλήρες πλαίσιο για την Ανάλυση Επιβίωσης. Αριστερή λογοκρισία προκύπτει όταν η έκβαση που μελετάται έχει ήδη συμβεί σε χρονικό σημείο πριν την έναρξη της παρατήρησης, χωρίς όμως να είναι γνωστός ο ακριβής χρόνος εμφάνισης της. Έστω ότι ο ερευνητής συλλέγει δεδομένα για τους συμμετέχοντες σε χρονική στιγμή c, και διαπιστώνεται ότι ορισμένα άτομα είχαν ήδη εμφανίσει το υπό μελέτη γεγονός. Σε αυτή την περίπτωση είναι γνωστό μόνο όταν ο χρόνος εμφάνισης του γεγονότος είναι μικρότερος από c, δηλαδή $T < c$, γεγονός που καθιστά τις παρατηρήσεις αυτές αριστερά λογοκριμένες. Αυτό συνεπάγεται ότι για τα άτομα των οποίων ο χρόνος εμφάνισης του γεγονότος είναι γνωστός (δηλαδή δεν έχουν υποστεί λογοκρισία τουλάχιστον μέχρι το χρόνο c), η ερμηνεία των δεδομένων βασίζεται στη p.d.f. $f(t;\boldsymbol{\theta})$ για το χρόνο *T.* Αντίθετα, για τις αριστερά λογοκριμένες παρατηρήσεις, όπου είναι γνωστό μόνο ότι $T < c$, χρησιμοποιείται η c.d.f. $F(c;\boldsymbol{\theta})$ του χρόνου *T,* η οποία εκφράζει την πιθανότητα το γεγονός να έχει ήδη συμβεί πριν το χρόνο παρατήρησης. Λαμβάνοντας υπόψη τις παραπάνω παραδοχές, η συνάρτηση πιθανοφάνειας για δεδομένα με αριστερή λογοκρισία στο χρόνο c προκύπτει ως εξής:

$$L(\boldsymbol{\theta}) = \prod_{i \in \Omega_1} f(t_i;\boldsymbol{\theta}) \times \prod_{j \in \Omega_2} F(c;\boldsymbol{\theta})$$

όπου με Ω_1 συμβολίζεται το σύνολο των δεδομένων χωρίς αριστερή λογοκρισία μέχρι το χρόνο c, ενώ Ω_2 αποτελεί το σύνολο των παρατηρήσεων που είναι αριστερά λογοκριμένες πριν το χρόνο c. Ισχύει ότι $\Omega = \Omega_1 \cup \Omega_2$ και $\Omega_1 \cap \Omega_2 = \emptyset$. Ένα ενδεικτικό παράδειγμα αριστερής λογοκρισίας μπορεί να προκύψει σε μελέτη που εξετάζει την επανεμφάνιση καρκίνου σε άτομα τα οποία είχαν υποβληθεί σε χειρουργική αφαίρεση όγκου. Εάν η μελέτη ξεκινά εκ των υστέρων, και κάποια άτομα έχουν ήδη εμφανίσει υποτροπή προτού εισέλθουν στη μελέτη, τότε οι παρατηρήσεις αυτές θεωρούνται αριστερά λογοκριμένες.

1.3.3 Λογοκρισία σε Διάστημα

Ο Lawless (2003) παρουσιάζει μία ακόμη σημαντική κατηγορία λογοκρισίας, η οποία μπορεί να θεωρηθεί ως γενίκευση των μορφών δεξιάς και αριστερής λογοκρισίας, είναι η λογοκρισία σε διάστημα (interval censoring). Στην περίπτωση αυτή, κάθε άτομο της μελέτης παρακολουθείται σε διακριτά χρονικά σημεία κατά τη διάρκεια της συνολικής περιόδου παρακολούθησης, τα οποία συμβολίζονται ως $0 = \alpha_0 < \alpha_1 < \cdots < \alpha_m < \infty$. Αν το άτομο δεν έχει εμφανίσει την υπό μελέτη έκβαση μέχρι το χρόνο α_i τότε επανεξετάζεται στο επόμενο χρονικό σημείο α_{i+1} προκειμένου να

διαπιστωθεί αν το γεγονός συνέβη εντός του διαστήματος $(a_i, a_{i+1}]$. Με αυτόν τον τρόπο, η χρονική στιγμή εμφάνισης της έκβασης T_i είναι γνωστό ότι ανήκει σε ένα διάστημα της μορφής $(U_i, V_i]$ με $U_i < T_i \leq V_i$. Αυτή η μορφή αποτελεί το γενικό τρόπο παρατήρησης λογοκριμένων χρόνων για την περίπτωση της λογοκρισίας σε διάστημα. Ειδικές περιπτώσεις της λογοκρισίας σε διάστημα αποτελούν η δεξιά λογοκρισία (όταν η έκβαση δεν έχει παρατηρηθεί μέχρι τον τελευταίο χρόνο παρακολούθησης $a_m = c$ και τεθεί $U_i = c$ και $V_i \to +\infty$) και η αριστερή λογοκρισία (όταν η έκβαση έχει ήδη συμβεί πριν από την πρώτη παρατήρηση, δηλαδή $U_i = 0$ και $V_i = c$ με $c \in (0, +\infty)$).

Ένα χαρακτηριστικό παράδειγμα λογοκρισίας σε διάστημα αποτελεί η περίπτωση των δεδομένων τρέχουσας κατάστασης (current status data). Στα δεδομένα αυτά, ο χρόνος εμφάνισης του γεγονότος είναι άγνωστος, ενώ η παρατήρηση γίνεται μόνο μία φορά, τη χρονική στιγμή C_i. Αν κατά την παρατήρηση διαπιστώνεται ότι το γεγονός έχει ήδη συμβεί, το χρονικό διάστημα του γεγονότος είναι $(0, C_i]$ (αριστερή λογοκρισία). Αν όχι, θεωρείται ότι $T_i > C_i$ (δεξιά λογοκρισία). Τέτοια δεδομένα συναντώνται σε ιατρικές μελέτες, όπως για παράδειγμα στην καρκινογένεση σε πειραματόζωα, όπου ο χρόνος εμφάνισης όγκου είναι αντικείμενο μελέτης, αλλά η διάγνωση μπορεί να γίνει μόνο κατά τη νεκροψία. Η συνάρτηση πιθανοφάνειας για λογοκριμένα δεδομένα σε διάστημα δίνεται από την ακόλουθη μορφή:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} [F(V_i; \boldsymbol{\theta}) - F(U_i; \boldsymbol{\theta})]$$

όπου $F(\cdot\,; \boldsymbol{\theta})$ είναι η c.d.f. του χρόνου T με παραμέτρους $\boldsymbol{\theta}$ και ισχύει ότι $F(0) = 0$. Η ποσότητα $F(V_i) - F(U_i)$ εκφράζει την πιθανότητα το γεγονός να έχει συμβεί εντός του διαστήματος $(U_i, V_i]$ για κάθε παρατήρηση i.

1.3.4 Ανεξάρτητη τυχαία λογοκρισία

Οι βασικότεροι τύποι λογοκρισίας παρουσιάστηκαν παραπάνω, μαζί με τις αντίστοιχες συναρτήσεις πιθανοφάνειας. Ωστόσο, ιδιαίτερη σημασία για την ανάλυση των δεδομένων επιβίωσης έχει η κατηγορία της ανεξάρτητης ή μη-πληροφοριακής τυχαίας λογοκρισίας (independent or non-informative censoring), η οποία αποτελεί βασική προϋπόθεση για την ορθότητα των στατιστικών μεθόδων που χρησιμοποιούνται. Ο Collett (2023) τονίζει τη σημασία της υπόθεσης περί ανεξαρτησίας μεταξύ του χρόνου επιβίωσης και του χρόνου λογοκρισίας, καθώς σε αντίθετη περίπτωση μεταβάλλεται ριζικά η ανάλυση. Σύμφωνα με αυτή την παραδοχή, θεωρείται ότι ο πραγματικός χρόνος επιβίωσης ενός ατόμου δεν εξαρτάται από το μηχανισμό ή το χρόνο λογοκρισίας C_i. Αυτό σημαίνει ότι, υπό δεδομένες τιμές των σχετικών προγνωστικών μεταβλητών (covariates), ένα λογοκριμένο άτομο μπορεί να θεωρηθεί αντιπροσωπευτικό της ομάδας ατόμων με παρόμοιο προφίλ, που έχουν παραμείνει υπό παρακολούθηση μέχρι τον ίδιο χρόνο. Ο Lawless (2003) εκφράζει την παραπάνω συνθήκη με τη χρήση των p.d.f. και s.f., όπως ορίστηκαν πρωτύτερα. Έστω ότι T_i και C_i είναι συνεχείς τυχαίες μεταβλητές που αναπαριστούν τον πραγματικό χρόνο επιβίωσης και το χρόνο λογοκρισίας του i −οστού ατόμου,

αντίστοιχα. Με $f(t)$ και $g(t)$ ορίζονται οι αντίστοιχες p.d.f. των T_i και C_i, ενώ $S(t)$ και $G(t)$ οι αντίστοιχες s.f.**.** Ορίζεται ακόμη η δείκτρια:

$$\delta_i = I(T_i \leq C_i)$$

που παίρνει τιμές 1 όταν το γεγονός (π.χ. θάνατος ή αποτυχία) παρατηρείται και 0 όταν το άτομο λογοκρίνεται. Τότε, η από κοινού p.d.f. του ζεύγους (T_i, C_i) δίνεται ως:

$$P(t_i = t, \delta_i) = \begin{cases} g(t)S(t), & \delta_i = 0 \\ f(t)G(t), & \delta_i = 1 \end{cases}$$

Η παραπάνω έκφραση μπορεί να διατυπωθεί συνοπτικά ως:

$$P(t_i = t, \delta_i) = [f(t)G(t)]^{\delta_i}[g(t)S(t)]^{1-\delta_i}$$

Υπό αυτή τη συνθήκη, η συνάρτηση πιθανοφάνειας για δείγμα μεγέθους n προκύπτει ως:

$$L(\cdot) = \prod_{i=1}^{n}[f(t_i)G(t_i)]^{\delta_i}[g(t_i)S(t_i)]^{1-\delta_i}$$

Ένα παράδειγμα ανεξάρτητης λογοκρισίας μπορεί να θεωρηθεί η περίπτωση όπου ένα άτομο αποχωρεί από τη μελέτη λόγω μετεγκατάστασης σε άλλη περιοχή, για λόγους που δεν σχετίζονται με την υγεία του ή την πιθανότητα εμφάνισης του υπό εξέταση γεγονότος (π.χ. υποτροπής). Επίσης, η λογοκρισία που προκύπτει όταν η ανάλυση των δεδομένων επιβίωσης πραγματοποιείται σε προκαθορισμένο χρονικό όριο (όπως είναι το τέλος της μελέτης) θεωρείται γενικά ανεξάρτητη, εφόσον όλοι οι συμμετέχοντες υπόκεινται στον ίδιο μηχανισμό λογοκρισίας ανεξαρτήτως του χρόνου επιβίωσης τους.

1.3.5 Εξαρτημένη λογοκρισία

Μία από τις πιο απαιτητικές και περίπλοκες μορφές λογοκρισίας, από στατιστικής απόψεως, αποτελεί η εξαρτημένη ή πληροφοριακή λογοκρισία (dependent or informative censoring), όπως περιγράφεται από τον Collett (2023). Στην περίπτωση αυτή, ο χρόνος μέχρι την εμφάνιση του μελετώμενου συμβάντος και ο χρόνος μέχρι τη λογοκρισία δεν είναι ανεξάρτητες μεταβλητές, αλλά σχετίζονται μεταξύ τους. Δηλαδή, ο μηχανισμός λογοκρισίας περιέχει πληροφορία για τον υποκείμενο χρόνο επιβίωσης. Τέτοιες περιπτώσεις απαιτούν εναλλακτικές στατιστικές προσεγγίσεις για την εξαγωγή αξιόπιστων συμπερασμάτων. Ωστόσο, η παρούσα εργασία δεν εξετάζει τέτοιες μεθόδους, καθώς υπερβαίνουν το θεωρητικό της πεδίο. Επιπλέον, η εξάρτηση μεταξύ της λογοκρισίας και χρόνου επιβίωσης συχνά δεν είναι απλή να ανιχνευθεί μόνο από τα δεδομένα, γεγονός που αυξάνει τον κίνδυνο μεροληπτικών εκτιμήσεων. Ωστόσο, σε ορισμένες περιπτώσεις, ο αναλυτής μπορεί να αξιοποιήσει στοιχεία από το σχεδιασμό της μελέτης ή την κλινική εμπειρία για να υποπτευθεί την πιθανή παρουσία εξαρτημένης λογοκρισίας, ακόμη κι αν αυτή δεν είναι άμεσα παρατηρήσιμη στα δεδομένα.

Η εγκυρότητα των κλασικών μεθόδων ανάλυσης επιβίωσης όπως είναι ο εκτιμητής της s.f. Kaplan – Meier, το μοντέλο αναλογικού κινδύνου Cox βασίζονται στην υπόθεση της ανεξαρτησίας. Σε περίπτωση παράβασης της (δηλαδή όταν η λογοκρισία είναι πληροφοριακή), οι εκτιμήσεις των s.f. ή των σχετικών κινδύνων ενδέχεται να είναι συστηματικά μεροληπτικές. Στις περιπτώσεις αυτές απαιτείται πιο σύνθετη μοντελοποίηση που ενσωματώνει το μηχανισμό λογοκρισίας. Ενδεικτικό παράδειγμα εξαρτημένης λογοκρισίας είναι οι περιπτώσεις μελετών όπου τα άτομα αποχωρούν από τη μελέτη λόγω επιδείνωσης της υγείας τους, ανεπιθύμητων παρενεργειών ή φόβου για συνέχιση της θεραπείας. Η αποχώρηση αυτή σχετίζεται με την πιθανότητα εκδήλωσης του μελετώμενου γεγονότος και συνεπώς παραβιάζει την υπόθεση ανεξαρτησίας.

1.4 Εκτίμηση βασικών συναρτήσεων επιβίωσης – Μη Παραμετρική Εκτίμηση

Στην υποενότητα 1.2 παρουσιάστηκαν οι πιο συνήθεις θεωρητικές κατανομές που μπορεί να ακολουθεί ο χρόνος επιβίωσης σε μία μελέτη. Ωστόσο, υπάρχουν περιπτώσεις όπου δεν είναι εφικτός ο σαφής προσδιορισμός της κατανομής από την οποία προέρχεται ο χρόνος επιβίωσης *T*. Σ' αυτές τις περιπτώσεις, η εκτίμηση των βασικών συναρτήσεων, όπως είναι η s.f. $S(t)$, η h.f. $h(t)$ και η c.h.f. $H(t)$, πραγματοποιείται με μη παραμετρικές μεθόδους. Στη Στατιστική Επιστήμη έχουν προταθεί διάφορες μη παραμετρικές προσεγγίσεις για την εκτίμηση αυτών των συναρτήσεων. Έστω δείγμα n παρατηρήσεων χωρίς λογοκρισία, με χρόνους επιβίωσης $t_1, t_2, \dots, t_n$. Η εμπειρική συνάρτηση επιβίωσης (empirical survival function**,** ESF) ορίζεται ως εξής:

$$\hat{S}(t) = \frac{\sum_{i=1}^{n} I(t_i > t)}{n}, t \geq 0 \qquad (1.12)$$

όπου $I(\cdot)$ είναι η δείκτρια που λαμβάνει την τιμή 1 όταν η συνθήκη ισχύει και 0 διαφορετικά. Αν όλοι οι παρατηρούμενοι χρόνοι είναι διακεκριμένοι, τότε η s.f. μειώνεται με βήμα $1/n$. Αντίστοιχα, σε περίπτωση που υπάρχουν d ταυτόσημοι χρόνοι επιβίωσης ίσοι με t, τότε η τιμή της s.f. μειώνεται με βήμα d/n αμέσως μετά το χρόνο t. Η γραφική αναπαράσταση αυτής της συνάρτησης είναι κλιμακωτή. Οι παρακάτω ορισμοί βασίζονται στο έργο του Lawless (2003).

1.4.1 Εκτίμηση της Συνάρτησης Επιβίωσης και Αθροιστικής Συνάρτησης Κινδύνου

Όταν υπάρχουν λογοκριμένες παρατηρήσεις (ιδίως με δεξιά λογοκρισία), η εκτίμηση της συνάρτησης επιβίωσης διαφοροποιείται. Ο εκτιμητής Kaplan – Meier (1958) αποτελεί τη βασική μέθοδο σε τέτοιες περιπτώσεις. Έστω ότι παρατηρούνται k διακριτοί χρόνοι επιβίωσης $t_1 < t_2 < \cdots < t_k$ στους οποίους συμβαίνει το υπό μελέτη γεγονός. Σε κάθε χρονική στιγμή t_j, μπορεί να εμφανιστούν πολλαπλά γεγονότα. Ορίζεται:

- d_j: αριθμός γεγονότων (π.χ. θανάτων) στο χρόνο t_j
- n_j: αριθμός ατόμων σε κίνδυνο ακριβώς πριν το χρόνο t_j

Οι υπόλοιπες $n-k$ παρατηρήσεις θεωρούνται λογοκριμένες, καθώς δεν έχει παρατηρηθεί το γεγονός έως το τέλος της παρακολούθησης.

Ο εκτιμητής Kaplan – Meier (KM) (ή Product – Limit Estimator) δίνεται από:

$$\hat{S}_{KM}(t) = \prod_{j:t_j<t} \frac{n_j - d_j}{n_j} \tag{1.13}$$

Σε περιπτώσεις όπου η λογοκρισία συμπίπτει χρονικά με έναν διακριτό χρόνο t_j, γίνεται η υπόθεση ότι το λογοκριμένο γεγονός συνέβη απειροελάχιστα μετά από το χρόνο t_j. Από τη μορφή του τύπου (1.13) προκύπτει ότι $\hat{S}_{KM}(0) = 1$, ενώ η συνάρτηση είναι αριστερά – συνεχής και κλιμακωτή. Σε περιβάλλον χωρίς λογοκρισία, ο Kaplan – Meier εκτιμητής συμπίπτει με την εμπειρική συνάρτηση (1.12). Ο Nelson (1969) και αργότερα ο Aalen (1972), πρότειναν έναν εκτιμητή για την αθροιστική συνάρτηση κινδύνου, ο οποίος ορίζεται ως:

$$\widehat{H}_{NA}(t) = \sum_{j:t_j \le t} \frac{d_j}{n_j}$$

Συνδυάζοντας τον παραπάνω τύπο με τη σχέση (1.6) προκύπτει ο εκτιμητής Altshuler γνωστός και ως εναλλακτική προσέγγιση της Kaplan – Meier:

$$\hat{S}_{NA}(t) = \prod_{j=1}^{k} e^{-\frac{d_j}{n_j}}$$

Σύμφωνα με το ανάπτυγμα Taylor:

$$e^{-x} = 1 - x + \frac{x^2}{2} - \frac{x^3}{6} + \cdots$$

προκύπτει ότι για δεδομένο $x > 0$, ισχύει $e^{-x} \geq 1 - x$. Συνεπώς, ο εκτιμητής Nelson – Aalen τείνει να δίνει μεγαλύτερες τιμές στη συνάρτηση επιβίωσης σε σχέση με τον εκτιμητή Kaplan – Meier. Ο Collett (2023) επισημαίνει ότι ο εκτιμητής Nelson – Aalen έχει καλύτερη απόδοση σε μικρά δείγματα, ενώ σε μεγαλύτερα δείγματα, οι δύο εκτιμητές συγκλίνουν (ιδιαίτερα στα αρχικά χρονικά διαστήματα).

1.4.2 Πίνακας Επιβίωσης

Οι πίνακες επιβίωσης αποτελούν μία από τις παλαιότερες μεθόδους ανάλυσης δεδομένων επιβίωσης, με εφαρμογές ήδη από τις αρχές του 20ου αιώνα, κυρίως στο πεδίο της Δημογραφίας. Η βασική ιδέα της μεθόδου έγκειται στη διαίρεση του συνολικού χρόνου παρακολούθησης σε $k+1$ διαστήματα της μορφής $I_j = [a_{j-1}, a_j)$ για $j = 1,2,\dots,k+1$ όπου $\alpha_0 = 0, \alpha_k = L$ (ο μέγιστος παρατηρούμενος χρόνος επιβίωσης) και $a_{k+1} = \infty$. Για κάθε παρατήρηση καταγράφεται είτε ο χρόνος του γεγονότος *T,* είτε ο χρόνος λογοκρισίας *C.* Οι παρατηρήσεις ομαδοποιούνται στα αντίστοιχα χρονικά διαστήματα και καταγράφονται τα παρακάτω μεγέθη:

- N_j: πλήθος ατόμων που βρίσκονται σε κίνδυνο στην αρχή του διαστήματος I_j (με $N_1 = n$)

- D_j: αριθμός γεγονότων (π.χ. θανάτων) που συνέβησαν εντός του διαστήματος I_j
- W_j: αριθμός λογοκριμένων παρατηρήσεων εντός του διαστήματος I_j

Οι τιμές αυτές συνδέονται μέσω της αναδρομικής σχέσης:

$$N_j = N_{j-1} - D_{j-1} - W_{j-1}, \qquad j = 2,3,\dots,k+1$$

Ορίζονται, επιπλέον οι ακόλουθες πιθανότητες:

$$p_j = \frac{P_j}{P_{j-1}} = \frac{N_j' - D_j}{N_j'}$$

όπου p_j είναι η εκτιμώμενη πιθανότητα επιβίωσης στο διάστημα I_j δεδομένου ότι το άτομο έχει επιβιώσει έως το I_{j-1} και $N_j' = N_j - 0.5\, W_j$ είναι ο διορθωμένος αριθμός ατόμων σε κίνδυνο, με προσαρμογή για τη λογοκρισία.

$$q_j = 1 - p_j = \frac{D_j}{N_j'}$$

όπου q_j είναι η εκτιμώμενη πιθανότητα εμφάνισης του γεγονότος εντός του διαστήματος I_j υπό την προϋπόθεση ότι το άτομο έχει επιβιώσει μέχρι την αρχή του διαστήματος και ισχύει ότι $q_{k+1} = 1$.

Με βάση τις παραπάνω πιθανότητες ορίζεται μία εκτίμηση της s.f., η οποία έχει παρόμοια μορφή με τον εκτιμητή Kaplan – Meier. Οι δύο μέθοδοι, ωστόσο, διαφέρουν σημαντικά καθώς, στους πίνακες επιβίωσης, ο χρόνος αναλύεται σε διακριτά χρονικά διαστήματα και οι λογοκριμένες παρατηρήσεις λαμβάνονται υπόψη (ως $W_j/2$) ενώ, στον εκτιμητή Kaplan – Meier, ο χρόνος διατηρείται συνεχής και οι λογοκριμένες παρατηρήσεις δεν επηρεάζουν τη χρονική στιγμή κατά την οποία συμβαίνει το γεγονός. Η τελική μορφή του εκτιμητή της s.f. από πίνακα επιβίωσης δίνεται από τον τύπο:

$$P_j = S(a_j) = \prod_{i=1}^{j} p_i = \prod_{i=1}^{j} \left(1 - \frac{D_j}{N_j'}\right)$$

1.4.3 Εκτίμηση Διακύμανσης των Μη Παραμετρικών Εκτιμητών

Πέρα από την εκτίμηση των βασικών συναρτήσεων που παρουσιάστηκαν στις υποενότητες 1.4.1 και 1.4.2, ιδιαίτερη σημασία για τη στατιστική ανάλυση αποκτά και η εκτίμηση της διακύμανσης των αντίστοιχων εκτιμητών. Η πληροφορία αυτή είναι απαραίτητη για τη διενέργεια στατιστικής συμπερασματολογίας, όπως για παράδειγμα για την κατασκευή διαστημάτων εμπιστοσύνης. Η διακύμανση της εκτίμησης της s.f. μέσω του Kaplan – Meier εκτιμητή υπολογίζεται μέσω του τύπου Greenwood:

$$\widehat{Var}\left(\hat{S}_{KM}(t)\right) = \hat{S}(t)^2 \sum_{j:t_j<t} \frac{d_j}{n_j(n_j - d_j)}$$

Για την εκτίμηση της c.h.f., η διακύμανση του εκτιμητή Nelson – Aalen δίνεται από τον τύπο:

$$\widehat{Var}\left(\hat{H}_{NA}(t)\right) = \sum_{j:t_j<t} \frac{d_j(n_j - d_j)}{n_j^3}$$

Σε πολλές περιπτώσεις χρησιμοποιείται και μία απλούστερη (συντηρητική) εκδοχή του παραπάνω τύπου:

$$\widehat{Var}\left(\hat{H}_{NA}(t)\right) = \sum_{j:t_j<t} \frac{d_j}{n_j^2}$$

Η εκτίμηση της διακύμανσης της s.f. που προκύπτει μέσω του Nelson – Aalen μπορεί στη συνέχεια να προσεγγιστεί μέσω του μετασχηματισμού (1.6) με χρήση της μεθόδου Δέλτα ή άλλων τεχνικών όπως είναι η μέθοδος bootstrap. Για την περίπτωση της εκτίμησης της s.f. με βάση τον πίνακα επιβίωσης, η διακύμανση του εκτιμητή $\hat{P}_j = \hat{S}(a_j)$ δίνεται από τη σχέση:

$$\widehat{Var}(\hat{P}_j) = \hat{P}_j^2 \sum_{i=1}^{j} \frac{\hat{q}_i}{N_i' \hat{p}_i}$$

όπου $N_i' = N_i - 0.5W_i$ και $\hat{p}_i, \hat{q}_i$ οι εκτιμώμενες πιθανότητες επιβίωσης και εμφάνισης του γεγονότος στο αντίστοιχο διάστημα I_j. Η τυπική απόκλιση (standard deviation) κάθε εκτιμητή υπολογίζεται ως η τετραγωνική ρίζα της αντίστοιχης διακύμανσης:

$$\widehat{s.d.}(\hat{\theta}) = \sqrt{\widehat{Var}(\hat{\theta})}$$

όπου $\hat{\theta}$ είναι κάποιος από τους εκτιμητές $\hat{S}_{KM}(t), \hat{H}_{NA}(t)$ ή $\hat{P}_j$.

1.5 Μοντέλα Παλινδρόμησης με εφαρμογή στην Ανάλυση Επιβίωσης

Στο πλαίσιο της στατιστικής ανάλυσης, συχνά εξετάζονται δεδομένα που αποτελούνται από n ανεξάρτητες παρατηρήσεις, καθεμία εκ των οποίων περιλαμβάνει μία μεταβλητή απόκρισης και p επεξηγηματικές μεταβλητές (covariates). Ο πρωταρχικός στόχος τέτοιου είδους αναλύσεων είναι η διερεύνηση της σχέσης μεταξύ της μεταβλητής απόκρισης και των επεξηγηματικών μεταβλητών καθώς και η διεξαγωγή προβλέψεων. Αρχικά, υλοποιείται μια περιγραφική στατιστική ανάλυση, με στόχο την κατανόηση της δομής του δείγματος, τον εντοπισμό πιθανών συσχετίσεων και την αποτύπωση των βασικών χαρακτηριστικών του υπό μελέτη πληθυσμού. Στη συνέχεια, εφαρμόζονται μέθοδοι στατιστικής συμπερασματολογίας για την εξαγωγή γενικεύσιμων συμπερασμάτων από το δείγμα προς τον πληθυσμό.

Στο πλαίσιο της Ανάλυσης Επιβίωσης, χρησιμοποιούνται εξειδικευμένες στατιστικές διαδικασίες για τη σύγκριση των s.f. μεταξύ δύο ή περισσότερων ομάδων. Οι πλέον διαδεδομένοι έλεγχοι περιλαμβάνουν τους: Log-rank, Gehan (ή Wilcoxon ή Breslow), Tarone – Ware, Peto – Peto, Modified Peto – Peto και Fleming – Harrington. Σε περιπτώσεις ύπαρξης περισσότερων από μίας ποιοτικών επεξηγηματικών μεταβλητών με διακριτά επίπεδα εφαρμόζονται στρωματοποιημένοι έλεγχοι. Σε αυτές τις περιπτώσεις, διατηρείται σταθερό ένα επίπεδο της μίας μεταβλητής (στρώμα), και εξετάζεται η ισότητα των s.f. στα επίπεδα της άλλης μεταβλητής για κάθε στρώμα της πρώτης ξεχωριστά. Επιπλέον, όταν η επεξηγηματική μεταβλητή είναι διατάξιμη (ordinal variable) είναι δυνατόν να εφαρμοστούν έλεγχοι τάσης (trend tests) όπως είναι ο έλεγχος τάσης Log – rank οι

οποίοι έχουν αυξημένη ισχύ όταν υπάρχει μονοτονική μεταβολή του κινδύνου ή του χρόνου επιβίωσης μεταξύ των διαδοχικών επιπέδων της μεταβλητής (π.χ. αυξανόμενος κίνδυνος ανάλογα με τη δόση θεραπείας ή το στάδιο νόσου).

Σε επόμενο στάδιο, αξιοποιούνται μοντέλα παλινδρόμησης, τα οποία αποσκοπούν στη διατύπωση μιας μαθηματικής σχέσης που συνδέει τη μεταβλητή απόκρισης με ένα σύνολο επεξηγηματικών μεταβλητών. Στην Ανάλυση Επιβίωσης, η μεταβλητή απόκρισης αντιστοιχεί στο χρόνο μέχρι την επέλευση ενός προκαθορισμένου γεγονότος (π.χ. θάνατος, υποτροπή νόσου, τεχνική αστοχία συστήματος). Οι επεξηγηματικές μεταβλητές οργανώνονται στον πίνακα $\boldsymbol{X}$ διαστάσεων $n \times p$, όπου κάθε γραμμή – διάνυσμα $\boldsymbol{X_i}$ περιέχει τις τιμές των p χαρακτηριστικών για την $i-$οστή παρατήρηση με $i = 1, \dots, n$ και κάθε στήλη – διάνυσμα $\boldsymbol{X_j}$ εκφράζει την $j-$οστή επεξηγηματική μεταβλητή μεγέθους n με $j = 1, \dots, p$. Τα χαρακτηριστικά αυτά μπορεί να σχετίζονται με δημογραφικές (π.χ. φύλο, ηλικία), θεραπευτικές (τύπος θεραπείας) ή άλλες κλινικές και λειτουργικές πληροφορίες. Στην παρούσα εργασία, οι επεξηγηματικές μεταβλητές θεωρούνται στατικές (δηλαδή λαμβάνονται κατά την έναρξη της μελέτης ($t = 0$)) και παραμένουν αμετάβλητες καθ' όλη τη διάρκεια της παρακολούθησης. Επιπλέον, γίνεται η υπόθεση ότι η λογοκρισία είναι δεξιά, μη πληροφοριακή και στατιστικά ανεξάρτητη από τις επεξηγηματικές μεταβλητές και το χρόνο επιβίωσης.

Η παρούσα ενότητα επικεντρώνεται σε δύο βασικές κατηγορίες μοντέλων παλινδρόμησης για δεδομένα επιβίωσης. Η πρώτη κατηγορία είναι τα ημιπαραμετρικά μοντέλα με χαρακτηριστικότερο το μοντέλο αναλογικού κινδύνου του Cox, το οποίο μοντελοποιεί την h.f. χωρίς να επιβάλλει κάποια παραδοχή για την κατανομή του χρόνου επιβίωσης *T*. Η προσέγγιση αυτή προσφέρει σημαντική ευελιξία και ευρεία εφαρμογή σε πραγματικά δεδομένα. Η δεύτερη κατηγορία είναι τα παραμετρικά μοντέλα κινδύνου, γνωστά και ως μοντέλα επιταχυνόμενου χρόνου αποτυχίας (AFT models). Τα μοντέλα αυτά εστιάζουν στη μοντελοποίηση του χρόνου επιβίωσης *T* ή του λογαρίθμου του ($\log T$), υποθέτοντας συγκεκριμένη κατανομή για το χρόνο επιβίωσης. Στην παρούσα ανάλυση εξετάζονται AFT μοντέλα με υποκείμενες κατανομές που αναλύθηκαν στην υποενότητα 1.2.3, όπως είναι οι Weibull, Extreme Value, Logistic, Log – Logistic, Normal, και Log – Normal.

Η θεμελιώδης διαφορά μεταξύ των δύο αυτών προσεγγίσεων έγκειται τόσο στην ποσοτική μεταβλητή που μοντελοποιείται, όσο και στις παραδοχές αναφορικά με την κατανομή του χρόνου επιβίωσης. Συγκεκριμένα, τα παραμετρικά μοντέλα παρέχουν άμεσες εκτιμήσεις των κατανομών και του χρόνου επιβίωσης, ενώ τα ημιπαραμετρικά μοντέλα, αν και λιγότερο εκφραστικά ως προς την πλήρη μορφή της κατανομής, χαρακτηρίζονται ως περισσότερο ευέλικτα και με περιορισμένες παραδοχές. Η θεωρητική βάση της παρούσας ενότητας βασίζεται κυρίως στο έργο του Collett (2023) το οποίο αποτελεί θεμελιώδες κείμενο στην περιοχή της Ανάλυσης Επιβίωσης.

1.5.1 Μοντέλο Αναλογικού Κινδύνου του Cox

1.5.1.1 Γενική μορφή μοντέλου αναλογικού κινδύνου

Το μοντέλο αναλογικού κινδύνου του Cox αποτελεί ένα από τα πιο θεμελιώδη και ευρέως χρησιμοποιούμενα μοντέλα παλινδρόμησης στην Ανάλυση Επιβίωσης. Ένα από τα σημαντικότερα χαρακτηριστικά του είναι ότι δεν απαιτεί την υπόθεση συγκεκριμένης κατανομής για το χρόνο εμφάνισης του υπό μελέτη συμβάντος *T,* αλλά εκτιμά τη συνάρτηση κινδύνου του χρόνου επιβίωσης *T,* με μη παραμετρικό τρόπο, την οποία και μοντελοποιεί. Έστω ότι εξετάζεται η αποτελεσματικότητα ενός νέου φαρμάκου (N) σε σύγκριση με μία ήδη υπάρχουσα φαρμακευτική αγωγή (S). Η ιδιότητα της αναλογίας κινδύνου PH προϋποθέτει ότι ισχύει η ακόλουθη σχέση:

$$h_N(t) = \psi\, h_S(t) \tag{1.14}$$

για κάθε μη αρνητική τιμή του χρόνου t, με τη σταθερά ψ να εκφράζει την αναλογία κινδύνου (relative hazard) μεταξύ των δύο θεραπειών. Η σχέση αυτή συνεπάγεται ότι οι αντίστοιχες συναρτήσεις επιβίωσης των δύο θεραπειών δεν τέμνονται. Η σταθερά ψ, που ονομάζεται σχετικός κίνδυνος (hazard ratio), εκφράζει την αναλογία του κινδύνου εμφάνισης του συμβάντος για τα άτομα που λαμβάνουν τη νέα θεραπεία σε σύγκριση με εκείνα που λαμβάνουν την υπάρχουσα. Σε περίπτωση που $\psi < 1$ τότε ο κίνδυνος είναι μικρότερος για τη νέα θεραπεία, ενώ στην αντίθετη περίπτωση όπου $\psi > 1$ ο κίνδυνος είναι μεγαλύτερος για τη νέα θεραπεία. Το μοντέλο μπορεί να γενικευθεί για περισσότερες επεξηγηματικές μεταβλητές. Αν προηγουμένως η θεραπεία θεωρούνταν ως μία δίτιμη μεταβλητή (0: υπάρχουσα, 1: νέα) , τώρα θεωρείται η ύπαρξη p επεξηγηματικών μεταβλητών. Για να εξασφαλιστεί ότι η τιμή του σχετικού κινδύνου είναι πάντα θετική, ορίζεται:

$$\psi = e^{\boldsymbol{\beta}^T \boldsymbol{x_i}},\ \mu\varepsilon\ \boldsymbol{\beta}^T \boldsymbol{x_i} = \sum_{j=1}^{p} \beta_j x_{ij}$$

Η συνάρτηση κινδύνου $h_S(t)$, που αντιστοιχεί στο άτομο με όλες τις επεξηγηματικές μεταβλητές ίσες με μηδέν, ορίζεται ως αναφορική συνάρτηση κινδύνου (baseline hazard function) και συμβολίζεται με $h_0(t)$. Ο τύπος (1.14) διαμορφώνεται στη γενικευμένη του μορφή ως εξής:

$$h_i(t) = e^{\boldsymbol{\beta^T x_i}}\, h_0(t) \tag{1.15}$$

Η σχέση αυτή ορίζει το ημιπαραμετρικό μοντέλο παλινδρόμησης του Cox (1972). Μία ισοδύναμη λογαριθμική μορφή της (1.15) είναι:

$$\log\left\{\frac{h_i(t)}{h_0(t)}\right\} = \boldsymbol{\beta}^T \boldsymbol{x_i} = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

Από το μοντέλο (1.15) προκύπτει ότι οι άγνωστες παράμετροι προς εκτίμηση είναι το διάνυσμα $\boldsymbol{\beta}$ το οποίο περιλαμβάνει τους συντελεστές των επεξηγηματικών μεταβλητών $X_1, X_2, \ldots, X_p$, καθώς και η αναφορική συνάρτηση κινδύνου $h_0(t)$.

1.5.1.2 Εκτίμηση των συντελεστών παλινδρόμησης υπό την απουσία δεσμών

Στο μοντέλο του Cox, οι συντελεστές $\boldsymbol{\beta}$ και η αναφορική συνάρτηση κινδύνου $h_0(t)$ μπορούν να εκτιμηθούν ξεχωριστά. Είναι δυνατό, δηλαδή, να εκτιμηθεί το πηλίκο $h_i(t)/h_0(t)$ χωρίς να απαιτείται εκτίμηση της ίδιας της συνάρτησης $h_0(t)$. Έστω ότι το προς ανάλυση σύνολο δεδομένων περιλαμβάνει n παρατηρήσεις της μορφής $(t_i, \delta_i, \boldsymbol{X_i})$. Θα παρουσιαστούν δύο διαφορετικές προσεγγίσεις: η πρώτη για την απουσία δεσμών και η δεύτερη για την παρουσία δεσμών. Για την κατανόηση των παρακάτω τύπων ορίζεται ως $R(t_j)$ το σύνολο κινδύνου, δηλαδή το σύνολο των ατόμων που δεν έχουν ακόμη παρουσιάσει το συμβάν και δεν έχουν λογοκριθεί μέχρι τη χρονική στιγμή t_j. Στην περίπτωση απουσίας δεσμών/ισοπαλιών (κάθε γεγονός συμβαίνει σε διαφορετική χρονική στιγμή), η μερική συνάρτηση πιθανοφάνειας (partial likelihood function), η οποία μεγιστοποιείται για την εκτίμηση των παραμέτρων $\boldsymbol{\beta}$, έχει τη μορφή:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left\{ \frac{\exp(\boldsymbol{\beta^T x_i})}{\sum_{k \in R(t_i)} \exp(\boldsymbol{\beta^T x_k})} \right\}^{\delta_i} \tag{1.16}$$

Για την εκτίμηση των συντελεστών $\boldsymbol{\beta}$, χρησιμοποιείται η μέθοδος Newton – Raphson, η οποία περιγράφεται αναλυτικά στο έργο του Collett (2023, p. 60).

1.5.1.3 Εκτίμηση των συντελεστών παλινδρόμησης υπό την παρουσία δεσμών

Το μοντέλο παλινδρόμησης του Cox προϋποθέτει ότι η συνάρτηση κινδύνου είναι συνεχής στο χρόνο. Η παρουσία δεσμών (πολλαπλά γεγονότα που συμβαίνουν την ίδια χρονική στιγμή) παραβιάζει αυτή την υπόθεση. Διάφοροι ερευνητές έχουν προτείνει τροποποιήσεις της μερικής συνάρτησης πιθανοφάνειας για την αντιμετώπιση των δεσμών όπως οι Kalbfleisch & Prentice (2002), ο Breslow (1974), ο Peto (1972), ο Efron (1977) και ο ίδιος ο Cox (1972). Ο τύπος των Kalbfleisch and Prentice (2002) έχει πολύπλοκη μορφή και γι' αυτό δεν αναλύεται εκτενέστερα. Ο Breslow (1974) πρότεινε μια απλούστερη μορφή της μερικής συνάρτησης πιθανοφάνειας, η οποία χρησιμοποιείται ευρέως στην πράξη:

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{r} \frac{\exp(\boldsymbol{\beta^T s_j})}{\{\sum_{l \in R(t_{(j)})} \exp(\boldsymbol{\beta^T x_l})\}^{d_j}} \tag{1.17}$$

όπου $s_j = \{s_{ij}\}_{p \times 1}$ είναι ένα διάνυσμα διαστάσεων $p \times 1$, με κάθε στοιχείο s_{ij} να είναι το άθροισμα της i – οστής μεταβλητής για όλα τα άτομα που εμφάνισαν το γεγονός στο χρόνο $t_{(j)}$:

$$s_{ij} = \sum_{k=1}^{d_j} x_{ijk}$$

και d_j ο αριθμός συμβάντων στη χρονική στιγμή t_j. Η παραπάνω μορφή είναι η επικρατέστερη σε πολλά στατιστικά πακέτα και γλώσσες προγραμματισμού. Ο Efron (1977) πρότεινε μια πιο ακριβή προσέγγιση, ειδικά όταν οι δεσμοί είναι συχνοί:

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{r} \frac{\exp(\boldsymbol{\beta}^T \boldsymbol{s_j})}{\prod_{k=1}^{d_j} \left[\sum_{l \in R(t_{(j)})} \exp(\boldsymbol{\beta}^T \boldsymbol{x_l}) - \frac{k-1}{d_j} \sum_{l \in D(t_{(j)})} \exp(\boldsymbol{\beta}^T \boldsymbol{x_l}) \right]} \quad (1.18)$$

όπου $D(t_{(j)})$ είναι το σύνολο των παρατηρήσεων που εμφάνισαν το γεγονός στο χρόνο $t_{(j)}$. Οι προσεγγίσεις των Breslow και Efron οδηγούν συχνά σε παρόμοια αποτελέσματα, εκτός αν οι δεσμοί είναι πολύ συχνοί ή μεγάλοι σε πλήθος, Τέλος, ο Cox (1972) πρότεινε ακόμη μια διατύπωση για την περίπτωση διακριτού χρόνου, η οποία βασίζεται σε διακριτοποίηση της χρονικής μεταβλητής ώστε να διατηρηθεί η υπόθεση της συνέχειας:

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{r} \frac{\exp(\boldsymbol{\beta}^T \boldsymbol{s_j})}{\sum_{l \in R(t_{(j)}; d_j)} \exp(\boldsymbol{\beta}^T \boldsymbol{s_l})} \quad (1.19)$$

όπου το σύνολο $R(t_{(j)}; d_j)$ περιλαμβάνει τις d_j παρατηρήσεις που επιλέγονται από το σύνολο κινδύνου $R(t_{(j)})$ στη χρονική στιγμή $(t_{(j)})$.
Μία διακριτή εκδοχή του μοντέλου παλινδρόμησης του Cox είναι:

$$\frac{h_i(t)}{1 - h_i(t)} = e^{\boldsymbol{\beta^T x_i}} \frac{h_0(t)}{1 - h_0(t)} \quad (1.20)$$

Ο Collett (2023) επισημαίνει ότι όταν οι χρονικές στιγμές είναι πολύ κοντινές (το πλάτος του διαστήματος είναι σχεδόν μηδέν), το μοντέλο (1.20) προσεγγίζει το συνεχές μοντέλο (1.15). Επιπλέον, αν θεωρηθεί $d_j = 1$ (απουσία δεσμών) τότε όλες οι σχέσεις (1.17), (1.18) και (1.19) ανάγονται στη μορφή της μερικής συνάρτησης πιθανοφάνειας (1.16).

1.5.1.4 Ερμηνεία συντελεστών παλινδρόμησης

Η ερμηνεία ενός συντελεστή β_j εξαρτάται κυρίως από τη φύση της αντίστοιχης επεξηγηματικής μεταβλητής. Για μια συνεχή μεταβλητή, από τη σχέση (1.15) και διατηρώντας σταθερές όλες τις υπόλοιπες μεταβλητές, προκύπτει ότι:

$$h_i(t) = e^{\beta_j x_{ij}} h_0(t) \quad (1.21)$$

Ο συντελεστής β_j ερμηνεύεται ως ο λογάριθμος της αναλογίας κινδύνου (log hazard ratio). Έστω η σύγκριση δύο ατόμων με τιμές *X* και *X+1* για την εξεταζόμενη μεταβλητή, τότε ισχύει:

$$\frac{h_j(t)}{h_i(t)} = \frac{e^{\beta_j(X+1)} h_0(t)}{e^{\beta_j X} h_0(t)} = \frac{e^{\beta_j(X+1)}}{e^{\beta_j X}} = e^{\beta_j}$$

Δηλαδή, η αναλογία κινδύνου μεταβάλλεται κατά e^{β_j} για κάθε μονάδα αύξησης της μεταβλητής *X*, με τις υπόλοιπες μεταβλητές σταθερές. Για μία ποιοτική (κατηγορική μεταβλητή με $m \geq 2$ κατηγορίες, εισάγονται $m-1$ βωβές μεταβλητές (dummy variables), συνεπώς και $m-1$ αντίστοιχα συντελεστές γ_j. Κάθε βωβή μεταβλητή I_j για $j = 2, \ldots, m$ λαμβάνει τιμή 1 όταν το άτομο ανήκει στην ομάδα j και 0 σε κάθε άλλη περίπτωση. Η πληροφορία για την 1η κατηγορία ενσωματώνεται στην

αναφορική συνάρτηση κινδύνου $h_0(t)$. Επομένως, κάθε συντελεστής γ_j εκφράζει τη μεταβολή του λογαρίθμου του λόγου κινδύνου μεταξύ της ομάδας j και της 1ης (κατηγορία αναφοράς).

1.5.1.5 Έλεγχος υποθέσεων – Διάστημα εμπιστοσύνης για τον j −συντελεστή παλινδρόμησης

Η σημειακή εκτίμηση των συντελεστών β_j αφορά το δείγμα και δεν επιτρέπει από μόνη της τη γενίκευση στο συνολικό πληθυσμό. Για το σκοπό αυτό, εφαρμόζονται έλεγχοι υποθέσεων ή υπολογίζονται διαστήματα εμπιστοσύνης. Για τους ελέγχους υποθέσεων, η γενική μορφή των υποθέσεων είναι

$$H_0: \beta = \beta_0 \text{ έναντι } H_1: \beta \neq \beta_0 \text{ (αμφίπλευρος έλεγχος)}$$
$$H_0: \beta \leq \beta_0 \text{ έναντι } H_1: \beta > \beta_0 \text{ (δεξιά μονόπλευρος έλεγχος)}$$
$$H_0: \beta \geq \beta_0 \text{ έναντι } H_1: \beta < \beta_0 \text{ (αριστερά μονόπλευρος έλεγχος)}$$

Ο συνηθέστερος είναι ο αμφίπλευρος έλεγχος για $\beta_0 = 0$. Η στατιστική συνάρτηση του ελέγχου Wald έχει τη μορφή:

$$T = \frac{\beta - \beta_0}{s.e.(\beta)}$$

Υπό τη μηδενική υπόθεση, η T ακολουθεί ασυμπτωτικά την Κανονική Κατανομή $N(0,1)$. Το $p - value$, για τον αμφίπλευρο έλεγχο, υπολογίζεται ως:

$$p - value_{\neq} = P(|T| > |T_{obs}| \mid T \sim N(0,1)) = 2(1 - \Phi(|T_{obs}|))$$

Ενώ για τους μονόπλευρους ελέγχους, δίνεται από τους τύπους:

$$p - value_{>} = P(T > T_{obs} \mid T \sim N(0,1)) = 1 - \Phi(T_{obs})$$

$$p - value_{<} = P(T < T_{obs} \mid T \sim N(0,1)) = \Phi(T_{obs})$$

όπου

$$\Phi(T_{obs}) = \Phi\left(\frac{\hat{\beta} - \beta_0}{s.e.(\hat{\beta})}\right)$$

και $\Phi(x)$ η c.d.f. της τυπικής κανονικής κατανομής $N(0,1)$ με

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du$$

Αν το $p - value$ είναι μικρότερο ή ίσο από το επίπεδο σημαντικότητας α και $\beta_0 = 0$, τότε απορρίπτεται η μηδενική υπόθεση. Αυτό σημαίνει ότι η αντίστοιχη επεξηγηματική μεταβλητή επηρεάζει το χρόνο επιβίωσης *T*. Σε αντίθετη περίπτωση, η μεταβλητή φαίνεται ότι δεν έχει σημαντική επίδραση στο χρόνο επιβίωσης *T*.

Έλεγχος Λόγου Πιθανοφανειών – ΕΛΠ (Likelihood Ratio Test - LRT)

Εναλλακτικά, χρησιμοποιείται ο Έλεγχος Λόγου Πιθανοφανειών (LRT), ειδικά όταν συγκρίνονται εμφωλευμένα μοντέλα (nested models). Δύο μοντέλα είναι εμφωλευμένα όταν έχουν την ίδια μεταβλητή απόκρισης και οι επεξηγηματικές μεταβλητές του ενός είναι υποσύνολο του άλλου. Έστω ότι υπάρχει ένα μοντέλο Α με

p επεξηγηματικές μεταβλητές και ένα μοντέλο Β με επιπλέον q επεξηγηματικές μεταβλητές (Α ⊂ Β) και οι ποσότητες $log(L_A)$, $log(L_B)$ οι λογάριθμοι των μερικών συναρτήσεων πιθανοφάνειας για τα μοντέλα Α και Β αντίστοιχα. Οι υποθέσεις του ΕΛΠ ορίζονται ως:

$$H_0: \beta_{p+1} = \beta_{p+2} = \cdots = \beta_{p+q} = 0 \text{ έναντι } H_1: \beta_j \neq 0 \text{ για κάποιο } j=p+1,\ldots,p+q$$

Η στατιστική συνάρτηση του ΕΛΠ ορίζεται ως

$$X^2 = -2\log(L_A/L_B)$$

Η ασυμπτωτική κατανομή της στατιστικής συνάρτησης X^2, υπό τη μηδενική υπόθεση, είναι η χ^2 κατανομή με q βαθμούς ελευθερίας. Αν ισχύει η σχέση $X^2_{obs} \geq \chi^2_{q,1-a}$ (όπου $\chi^2_{q,1-a}$ είναι το $\alpha\% -$ άνω ποσοστιαίο σημείο της κατανομής χ^2_q), το «μεγαλύτερο» μοντέλο (Β) προσφέρει στατιστικά σημαντική βελτίωση. Αν όχι, προτιμάται το απλούστερο μοντέλο (Α).

Διαστήματα Εμπιστοσύνης

Το $100(1-\alpha)\%$ διάστημα εμπιστοσύνης για το συντελεστή β_j είναι

$$\widehat{\beta_j} \pm z_{a/2}\, s.e.(\widehat{\beta_j})$$

Το διάστημα εμπιστοσύνης απεικονίζει ένα εύρος τιμών μέσα στο οποίο μπορεί να βρίσκεται η πραγματική τιμή του συντελεστή β_j, με περιθώριο σφάλματος $\alpha\%$. Αν το διάστημα δεν περιλαμβάνει την τιμή 0 (ή την υποθετική τιμή β_0) αυτό σημαίνει ότι η συγκεκριμένη επεξηγηματική μεταβλητή επηρεάζει το χρόνο επιβίωσης *T*. Αν το διάστημα περιλαμβάνει την τιμή 0, τότε δεν υπάρχουν αρκετές ενδείξεις ώστε να υποστηρίξουν τυχόν επίδραση της $j-$μεταβλητής στο χρόνο επιβίωσης σε επίπεδο σημαντικότητας $\alpha\%$. Τέλος, η παραπάνω στατιστική συμπερασματολογία για τα β_j μπορεί αντίστοιχα να εφαρμοσθεί και για τους συντελεστές γ_j.

1.5.2 Μοντέλα Επιταχυνόμενου Χρόνου Ζωής

1.5.2.1 Γενική μορφή AFT μοντέλων

Η πιο απλή περίπτωση εφαρμογής μοντέλων επιταχυνόμενου χρόνου ζωής αφορά τη σύγκριση δύο ομάδων, όπως για παράδειγμα μια ομάδα ατόμων που έλαβε ένα νέο φάρμακο και μια άλλη που έλαβε το υπάρχον (συμβατικό) φάρμακο. Ο ερευνητής διαθέτει τις παρατηρήσεις του χρόνου επιβίωσης για κάθε άτομο, περιλαμβανομένης ενδεχόμενης λογοκρισίας καθώς και μια δυαδική επεξηγηματική μεταβλητή που δηλώνει την ομάδα στην οποία ανήκει το άτομο (π.χ. 0=παλιά θεραπεία, 1=νέα θεραπεία). Ο βασικός στόχος είναι να διερευνηθεί αν υπάρχει μια πολλαπλασιαστική σχέση μεταξύ των συναρτήσεων επιβίωσης των δύο ομάδων, η οποία περιγράφεται μαθηματικά ως:

$$S_{new}(t) = S_{old}(t/\varphi)$$

όπου $S_{new}(t)$ είναι η συνάρτηση επιβίωσης για την ομάδα που έλαβε τη νέα θεραπεία ενώ S_{old} είναι η συνάρτηση επιβίωσης για την ομάδα ελέγχου (υπάρχουσα θεραπεία)

και η παράμετρος φ, ένας άγνωστος θετικός παράγοντας επιτάχυνσης (accelerated factor), ο οποίος εκτιμάται από τα δεδομένα.

Η παράμετρος φ αποτελεί το βασικό μέγεθος ερμηνείας στο AFT μοντέλο. Αν $\varphi > 1$, ο χρόνος επιβίωσης είναι μεγαλύτερος στην ομάδα που έλαβε τη νέα θεραπεία, κάτι που υποδηλώνει θετική επίδραση (βελτίωση). Ενώ αν $\varphi < 1$, τότε ο χρόνος επιβίωσης είναι μικρότερος και επομένως υπάρχει αρνητική επίδραση (επιδείνωση). Στην ειδική περίπτωση όπου $\varphi = 1$ τότε οι δύο ομάδες εμφανίζουν παρόμοια κατανομή επιβίωσης χωρίς ενδείξεις για διαφορά. Η ερμηνεία της παραμέτρου φ εξαρτάται από το αν το υπό μελέτη γεγονός είναι επιθυμητό (π.χ. θεραπεία μιας νόσου, βελτίωση κάποιας κατάστασης) ή ανεπιθύμητο (π.χ. θάνατος, αποτυχία θεραπείας), τότε μεγαλύτεροι χρόνοι επιβίωσης θεωρούνται ευνοϊκοί και επομένως $\varphi > 1$ δηλώνει βελτιωτική επίδραση. Αντίθετα, όταν το γεγονός είναι επιθυμητό (π.χ. αποθεραπεία), η ερμηνεία μπορεί να αντιστρέφεται. Συνεπώς, η κατεύθυνση της ερμηνείας πρέπει να αξιολογείται βάσει του συγκεκριμένου πλαισίου εφαρμογής. Από τις σχέσεις (1.4) και (1.21) μπορούν να προκύψουν οι αντίστοιχες σχέσεις για τις p.d.f και h.f.:

$$f_{new}(t) = \varphi^{-1} f_{old}(t/\varphi)$$
$$h_{new}(t) = \varphi^{-1} h_{old}(t/\varphi) \quad (1.22)$$

Έστω ότι X είναι μία δείκτρια επεξηγηματική συνάρτηση με $x_i = 0$ όταν το $i-$ άτομο λαμβάνει την παλιά θεραπεία ενώ $x_i = 1$ όταν το $i-$ άτομο λαμβάνει τη νέα θεραπεία. Τότε, η γενίκευση της (1.22) για κάθε $i-$ άτομο είναι:

$$h_i(t) = \varphi^{-x_i} h_0(t/\varphi^{x_i}) \quad (1.23)$$

Ωστόσο, για να διασφαλιστεί ότι η παράμετρος φ είναι θετική – όπως απαιτείται – εκφράζεται ως $\varphi = e^{\alpha}$ όπου $\alpha \in \mathbb{R}$. Αντικαθιστώντας στη σχέση (1.23) τη νέα μορφή της παραμέτρου φ, ο τύπος (1.23) γίνεται:

$$h_i(t) = e^{-\alpha x_i} h_0(t/e^{\alpha x_i})$$

Η παραπάνω εξίσωση αποτελεί τη βασική μορφή της συνάρτησης κινδύνου στα AFT μοντέλα στην περίπτωση μιας δυαδικής επεξηγηματικής μεταβλητής. Η παράμετρος α εκτιμάται μέσω μεθόδων μεγίστης πιθανοφάνειας. Όλα τα παραπάνω περιγράφουν την απλή και ενδεικτική περίπτωση όπου το AFT μοντέλο εφαρμόζεται σε μία μόνο δυαδική επεξηγηματική μεταβλητή. Στην πράξη, ο ερευνητής συνήθως διαθέτει ένα πλήθος επεξηγηματικών μεταβλητών (έστω p μεταβλητές) οι οποίες ενδέχεται να σχετίζονται με το χρόνο επιβίωσης. Σε αυτή την περίπτωση, στόχος είναι η μοντελοποίηση του T ως συνάρτηση ενός γραμμικού συνδυασμού των επεξηγηματικών μεταβλητών. Η γενικευμένη μορφή της συνάρτησης κινδύνου για το $i-$οστό άτομο σε ένα AFT μοντέλο δίνεται από την εξίσωση:

$$h_i(t) = e^{-\eta_i} h_0(t/e^{\eta_i}) \quad (1.24)$$

όπου $h_i(t)$ είναι η h.f. του $i-$οστού άτομο, $h_0(t)$ είναι η αναφορική συνάρτηση κινδύνου και $\eta_i = \boldsymbol{a}^T \boldsymbol{x_i} = \alpha_1 x_{i1} + \alpha_2 x_{i2} + \cdots + \alpha_p x_{ip}$ είναι ο γραμμικός

συνδυασμός του μοντέλου με x_{ij} να δηλώνει την τιμή της j – επεξηγηματικής μεταβλητής για το i −οστό άτομο και α_j τις αντίστοιχες παραμέτρους προς εκτίμηση.

Η παραπάνω μορφή αποτελεί τη βάση για τη μεθοδολογική γενίκευση των AFT μοντέλων και επιτρέπει την ανάλυση της επίδρασης πολλών παραγόντων (π.χ. φύλο, ηλικία, είδος θεραπείας, βιοχημικοί δείκτες κ.λπ.) στο χρόνο επιβίωσης. Οι παράμετροι α_j είναι οι άγνωστες παράμετροι των p επεξηγηματικών μεταβλητών και εκτιμώνται μέσω μεθόδων μεγίστης πιθανοφάνειας, ανάλογα με την κατανομή που υιοθετείται για το χρόνο επιβίωσης *T*. Το μοντέλο της εξίσωσης (1.24) μοντελοποιεί άμεσα τη συνάρτηση κινδύνου του χρόνου επιβίωσης *T*. Ωστόσο, στα AFT μοντέλα προτείνεται συχνά μία λογαριθμική (log – linear) αναδιατύπωση, η οποία λαμβάνει τη μορφή:

$$\log(T_i) = \mu + \boldsymbol{a}^T \boldsymbol{x_i} + \sigma \cdot \varepsilon_i \tag{1.25}$$

Στο μοντέλο (1.25), το διάνυσμα $\boldsymbol{a}$ περιλαμβάνει τους συντελεστές των επεξηγηματικών μεταβλητών, ενώ οι παράμετροι μ και σ αντιστοιχούν στο σταθερό όρο (intercept) και την παράμετρο κλίμακας (scale parameter) αντιστοίχως. Η ποσότητα ε_i εκφράζει το σφάλμα (error term), την τυχαία απόκλιση της τιμής $\log(T_i)$ από την αναμενόμενη τιμή που προσδιορίζεται από το γραμμικό μέρος του μοντέλου. Η παράμετρος a_i εκφράζει την επίδραση της αντίστοιχης επεξηγηματικής μεταβλητής στον λογάριθμο του χρόνου επιβίωσης. Θετική τιμή υποδηλώνει ότι η αύξηση της μεταβλητής (ή η μετάβαση σε άλλη κατηγορία, στην περίπτωση ποιοτικών μεταβλητών) σχετίζεται με αύξηση του χρόνου επιβίωσης, ενώ αρνητική τιμή δηλώνει το αντίθετο.

Με βάση τη μορφή του μοντέλου (1.25), η συνάρτηση επιβίωσης και η συνάρτηση κινδύνου για το $i-$ άτομο λαμβάνουν τις εξής μορφές:

$$\begin{gathered} S_i(t) = S_0\left\{\frac{t}{\exp(\boldsymbol{a^T x_i})}\right\} = S_{\varepsilon_i}\left(\frac{\log(t) - \mu - \boldsymbol{a^T x_i}}{\sigma}\right) \\ h_i(t) = e^{-\boldsymbol{a^T x_i}} h_0(t/e^{\boldsymbol{a^T x_i}}) \end{gathered} \tag{1.26}$$

Στις παραπάνω εξισώσεις, $S_0(t)$ και $h_0(t)$ είναι η αναφορική συνάρτηση επιβίωσης (baseline survival function) και η αναφορική συνάρτηση κινδύνου (baseline hazard function) αντίστοιχα, οι οποίες αναφέρονται σε άτομο με $\boldsymbol{x} = \boldsymbol{0}$ (όλες οι επεξηγηματικές μεταβλητές λαμβάνουν μηδενικές τιμές). Επιπλέον, $S_{\varepsilon_i}(\cdot)$ είναι η συνάρτηση επιβίωσης της τυχαίας μεταβλητής ε_i όπου εξαρτάται από την υπόθεση που γίνεται για την κατανομή της (π.χ. logistic, normal, extreme value).

1.5.2.2 Weibull & Extreme Value AFT μοντέλο

Υποθέτοντας ότι ο χρόνος επιβίωσης *T* ακολουθεί την κατανομή Weibull με παραμέτρους κλίμακας λ και σχήματος γ, η αναφορική συνάρτηση κινδύνου λαμβάνει τη μορφή:

$$h_0(t) = \lambda\gamma t^{\gamma-1} \tag{1.27}$$

Η h.f. για το $i-$ άτομο, βάσει του τύπου (1.24), δίνεται από τη σχέση:

$$h_i(t) = e^{-\eta_i}\lambda\gamma(e^{-\eta_i}t)^{\gamma-1} = (e^{-\eta_i})^{\gamma}\lambda\gamma t^{\gamma-1} \quad (1.28)$$

Από τον τύπο (1.28) προκύπτει ότι ο χρόνος επιβίωσης

$$T_i \sim Weib(e^{-\gamma\eta_i}, \gamma) \equiv Weib(e^{-\gamma \boldsymbol{a}^T \boldsymbol{x_i}}, \gamma)$$

Οι παράμετροι του μοντέλου εκτιμώνται μέσω μεγιστοποίησης της συνάρτησης πιθανοφάνειας, όπως ορίζεται στον τύπο (1.11), αντικαθιστώντας την p.d.f. και s.f. της Weibull κατανομής με παραμέτρους $e^{-\gamma \boldsymbol{a}^T \boldsymbol{x_i}}$ και γ. Στη λογαριθμική εκδοχή του AFT μοντέλου, όταν ο χρόνος T_i ακολουθεί Weibull κατανομή, η στοχαστική μεταβλητή ε_i ακολουθεί κατανομή Ακραίων Τιμών τύπου I (Extreme Value Type I), γνωστή και ως Gumbel κατανομή. Πρόκειται για ασύμμετρη κατανομή με μέση τιμή περίπου 0.5772 (σταθερά του Euler) και διακύμανση $\pi^2/6$. Σχετικά χαρακτηριστικά της κατανομής αυτής παρουσιάζονται στην υποενότητα 1.2.3, θέτοντας τις παραμέτρους $u = 0$ και $\beta = 1$. Για να αποδειχθεί ότι ο χρόνος επιβίωσης

$$T_i = \mathrm{e}^{\mu + \boldsymbol{a}^T \boldsymbol{x_i} + \sigma\, \varepsilon_i}$$

ακολουθεί Weibull κατανομή, γίνεται χρήση της s.f. της κατανομής $EV(0,1)$ από την οποία προκύπτει:

$$S_i(t) = \exp\left\{-exp\left(\frac{log(t) - \mu - \boldsymbol{a}^T \boldsymbol{x_i}}{\sigma}\right)\right\} = \exp(-\lambda_i t^{1/\sigma}) \quad (1.29)$$

όπου

$$\lambda_i = \exp\left(-\frac{(\mu + \boldsymbol{a}^T \boldsymbol{x_i})}{\sigma}\right)$$

αντιστοιχεί στην s.f. της Weibull κατανομής με παραμέτρους κλίμακας λ_i και σχήματος $1/\sigma$. Η c.h.f. για το μοντέλο Weibull – AFT μοντέλο διαμορφώνεται, βάσει της σχέσης (1.10) ως εξής:

$$H_i(t) = -\log\left(S_i(t)\right) = exp\left(\frac{log(t) - \mu - \boldsymbol{a}^T \boldsymbol{x_i}}{\sigma}\right) = \lambda_i t^{1/\sigma}$$

Η h.f., ως παράγωγος της $H_i(t)$ ως προς t, προκύπτει:

$$h_i(t) = \frac{1}{\sigma t} exp\left(\frac{log(t) - \mu - \boldsymbol{a}^T \boldsymbol{x_i}}{\sigma}\right)$$

Η Weibull κατανομή έχει το πλεονέκτημα ότι υποστηρίζει τόσο την επιταχυνόμενη ιδιότητα χρόνου αστοχίας (AFT) όσο και την αναλογική ιδιότητα κινδύνου (PH). Εφόσον η αναφορική συνάρτηση κινδύνου ακολουθεί Weibull κατανομή με παραμέτρους λ, γ (τύπος 1.27) και το εξεταζόμενο μοντέλο έχει τη μορφή αναλογικού κινδύνου (1.15), η συνάρτηση κινδύνου του i −οστού ατόμου είναι:

$$h_i(t) = \exp(\boldsymbol{\beta}^T \boldsymbol{x_i})\, \lambda\gamma^{t-1} h_i(t) = \exp(\boldsymbol{\beta}^T \boldsymbol{x_i})\, \lambda\gamma^{t-1}$$

που συνεπάγεται ότι $T_i \sim Weib(\lambda \exp(\boldsymbol{\beta}^T \boldsymbol{x_i}), \gamma)$ και η αντίστοιχη s.f. έχει τη μορφή:

$$S_i(t) = exp\{-exp(\boldsymbol{\beta}^T \boldsymbol{x_i})\lambda t^{\gamma}\} \quad (1.30)$$

Σε αυτά τα μοντέλα, οι επεξηγηματικές μεταβλητές επιδρούν αποκλειστικά στην παράμετρο λ, γεγονός που επηρεάζει το πόσο «γρήγορα» ή «αργά» συμβαίνει το υπό μελέτη γεγονός χωρίς να αλλάζει η μορφή της h.f.**.** Ο Collett (2023) επισημαίνει μια σημαντική σχέση μεταξύ των AFT και PH μοντέλων· αν οι συντελεστές $\boldsymbol{\alpha}$ του AFT μοντέλου πολλαπλασιαστούν με το $(-\gamma)$, προκύπτουν οι αντίστοιχοι συντελεστές $\boldsymbol{\beta}$ του μοντέλου αναλογικού κινδύνου (PH). Τέλος, η συνάρτηση επιβίωσης του μοντέλου PH (1.30) μπορεί να αναδιατυπωθεί ως:

$$S_i(t) = \exp\{-exp(\gamma\, log(t) + log(\lambda) + \boldsymbol{\beta}^T \boldsymbol{x_i})\}$$

Από τη μορφή αυτή προκύπτει μια άμεση σύνδεση μεταξύ των δύο προαναφερθέντων μοντέλων. Συγκρίνοντας τις σχέσεις (1.29) και (1.30), προκύπτουν οι ακόλουθες αντιστοιχίες:

$$\lambda = \exp(-\mu/\sigma), \quad \gamma = \sigma^{-1}, \quad \beta_j = -a_j/\sigma \quad για\, j = 1,2,\dots,p$$

1.5.2.3 Αναλογικό μοντέλο κινδύνου Gompertz

Σε συνέχεια της παρουσίασης του μοντέλου Weibull – AFT, το οποίο εντάσσεται τόσο στα AFT όσο και στα PH μοντέλα, παρουσιάζεται το αναλογικό μοντέλο κινδύνου Gompertz. Το μοντέλο αυτό βρίσκει σημαντικές εφαρμογές στη Δημογραφία και στις Επιστήμες Υγείας, καθώς η μορφή της h.f. αποδίδει ρεαλιστικά τη σταδιακή και εκθετική αύξηση του κινδύνου με το χρόνο, φαινόμενο συχνό σε πληθυσμιακές μελέτες και βιολογικά δεδομένα. Η βασική συνάρτηση κινδύνου, υπό την παραδοχή ότι ο χρόνος επιβίωσης ακολουθεί την κατανομή Gompertz με παραμέτρους $\lambda > 0$ και $\theta \in \mathbb{R}$, δίνεται από τη σχέση:

$$h_0(t) = \lambda e^{\theta t}$$

Η h.f. για το i −οστό άτομο, στο πλαίσιο του μοντέλου PH, προκύπτει ως:

$$h_i(t) = \lambda\, exp(\boldsymbol{\beta}^T \boldsymbol{x_i})\, exp(\theta t)$$

όπου $\boldsymbol{\beta}$ είναι το διάνυσμα των συντελεστών του μοντέλου για τις επεξηγηματικές μεταβλητές $\boldsymbol{X}$. Η κατανομή του χρόνου επιβίωσης T_i που προκύπτει από το μοντέλο είναι επίσης Gompertz, με παραμέτρους $\lambda\, exp(\boldsymbol{\beta}^T \boldsymbol{x_i})$ και θ. Σε αντίθεση με το Weibull – AFT μοντέλο, το μοντέλο Gompertz δεν υποστηρίζει την υπόθεση επιταχυνόμενου χρόνου αστοχίας (AFT) αλλά εντάσσεται αποκλειστικά στο πλαίσιο των αναλογικών μοντέλων κινδύνου (PH), καθώς ο λόγος κινδύνου μεταξύ δύο ατόμων παραμένει σταθερός σε κάθε χρονική στιγμή. Η εκτίμηση των άγνωστων παραμέτρων γίνεται μέσω μεγιστοποίησης της κατάλληλα διαμορφωμένης συνάρτησης πιθανοφάνειας, όπως αυτή δίνεται στον τύπο (1.11).

Συνοψίζοντας, το μοντέλο Gompertz προσφέρει μια παραμετρική προσέγγιση στο πλαίσιο αναλογικού κινδύνου, κατάλληλη για περιπτώσεις όπου ο κίνδυνος αυξάνεται εκθετικά με το χρόνο αν η παράμετρος $\theta > 0$, όπως στη μελέτη της γήρανσης. Αντίθετα, ο κίνδυνος μειώνεται εκθετικά όταν $\theta < 0$, ενώ στην ειδική περίπτωση του $\theta = 0$, το μοντέλο ισοδυναμεί με την εκθετική κατανομή. Στη συνέχεια, η ανάλυση επικεντρώνεται ξανά σε AFT μοντέλα με χρήση κατανομών όπως η Normal, Log-

Normal, Logistic και Log-Logistic που προσφέρουν μεγαλύτερη ευελιξία στη μορφή της s.f..

1.5.2.4 Log – Logistic & Logistic AFT μοντέλο

Στην παρούσα υποενότητα μελετάται το παραμετρικό AFT μοντέλο όπου ο χρόνος επιβίωσης T ακολουθεί Λογαριθμολογιστική κατανομή. Η βασική μορφή της κατανομής έχει ήδη παρουσιαστεί στην υποενότητα 1.2.3, ενώ εδώ χρησιμοποιείται μια εναλλακτική ισοδύναμη παραμετροποίηση. Η αναφορική συνάρτηση κινδύνου με παραμέτρους $\theta \in \mathbb{R}, \kappa > 0$ ορίζεται ως:

$$h_0(t) = \frac{e^{\theta}\ \kappa\ t^{k-1}}{1 + e^{\theta} t^{k}} \tag{1.31}$$

Υπό την υπόθεση ότι στη συνάρτηση κινδύνου της υποενότητας 1.2.3 τεθούν $\beta = \kappa$ και $\alpha = e^{-\theta/\kappa}$ προκύπτει ακριβώς η παραπάνω μορφή. Η γενίκευση για το i −οστό άτομο, σύμφωνα με τον τύπο (1.24), διαμορφώνεται ως:

$$h_i(t) = \frac{e^{\theta - \kappa \boldsymbol{a}^{\boldsymbol{T}} \boldsymbol{x_i}}\ \kappa\ t^{\kappa-1}}{1 + e^{\theta - \kappa \boldsymbol{a}^{\boldsymbol{T}} \boldsymbol{x_i}}\ t^{\kappa}} \tag{1.32}$$

Αντίστοιχα, η s.f. γράφεται ως:

$$S_i(t) = \frac{1}{1 + e^{\theta - \kappa \boldsymbol{a}^{\boldsymbol{T}} \boldsymbol{x_i}}\ t^{\kappa}} \tag{1.33}$$

Ο χρόνος επιβίωσης T_i, του οποίου η h.f. δίνεται από την (1.32), ακολουθεί την κατανομή $Log - Logistic\,(\theta - \kappa \boldsymbol{a}^T \boldsymbol{x_i}\ ,\ \kappa)$. Η Log – Logistic κατανομή χαρακτηρίζεται από την επιταχυνόμενη ιδιότητα χρόνου αστοχίας (AFT), ενώ δεν ικανοποιεί την ιδιότητα του αναλογικού κινδύνου (non – proportional hazards). Στο πλαίσιο της εναλλακτικής (λογαριθμικής) μορφής των AFT μοντέλων (τύπος 1.25), αν η στοχαστική μεταβλητή ε_i ακολουθεί τη Λογιστική Κατανομή με μέση τιμή 0 και διακύμανση $\pi^2/3$, τότε η s.f. του i −οστού ατόμου είναι:

$$S_i(t) = \left\{1 + \exp\left(\frac{\log(t) - \mu - \boldsymbol{a}^T \boldsymbol{x_i}}{\sigma}\right)\right\}^{-1}$$

Η παραπάνω λογαριθμική αναπαράσταση του AFT μοντέλου μπορεί να ταυτιστεί με την s.f. (1.33) θέτοντας $\theta = -\mu/\sigma$ και $\kappa = \sigma^{-1}$. Για την εκτίμηση της c.h.f. και στα δύο μοντέλα, εφαρμόζεται η σχέση (1.10). Συγκεκριμένα, για το μοντέλο με αναφορική συνάρτηση κινδύνου (1.31), προκύπτει:

$$H_i(t) = \log\left(1 + e^{\theta - \kappa \boldsymbol{a}^{\boldsymbol{T}} \boldsymbol{x_i}} t^{\kappa}\right)$$

Για τη λογαριθμική παραλλαγή του μοντέλου ισχύει:

$$H_i(t) = \log\left(1 + exp\left(\frac{log(t) - \mu - \boldsymbol{a}^T \boldsymbol{x_i}}{\sigma}\right)\right)$$

1.5.2.5 Log – Normal & Normal AFT μοντέλο

Υποθέτοντας ότι ο χρόνος επιβίωσης T ακολουθεί Λογαριθμοκανονική Κατανομή, η αναφορική συνάρτηση επιβίωσης δίνεται από τον τύπο:

$$S_0(t) = 1 - \Phi\left(\frac{\log(t) - \mu}{\sigma}\right) \tag{1.34}$$

όπου μ, σ^2 είναι οι παράμετροι της κατανομής και $\Phi(\cdot)$ η c.d.f. της τυπικής κανονικής κατανομής $N(0,1)$. Με βάση τους τύπους (1.26) και (1.34), το AFT μοντέλο γράφεται ως:

$$S_i(t) = S_0\left(t\, e^{-\boldsymbol{\alpha}^T \boldsymbol{x_i}}\right)$$

Ισοδύναμα, η s.f. του i −οστού ατόμου μπορεί να εκφραστεί ως:

$$S_i(t) = 1 - \Phi\left(\frac{\log(t) - \boldsymbol{\alpha}^T \boldsymbol{x_i} - \mu}{\sigma}\right)$$

Συνεπώς, ο χρόνος επιβίωσης T ακολουθεί Λογαριθμοκανονική Κατανομή με μέση τιμή $\boldsymbol{\alpha}^T \boldsymbol{x_i} + \mu$ και διακύμανση σ^2. Όπως και στις προηγούμενες περιπτώσεις, η Log – Normal κατανομή διαθέτει την επιταχυνόμενη ιδιότητα χρόνου αστοχίας. Για τη λογαριθμική μορφή του παραμετρικού AFT μοντέλου, η ποσότητα $log(T_i)$ ακολουθεί Κανονική Κατανομή με παραμέτρους μ, σ^2 γεγονός που συνεπάγεται ότι η στοχαστική μεταβλητή ε_i ακολουθεί την τυπική κανονική κατανομή $N(0,1)$.

Η s.f. γράφεται ως:

$$S_i(t) = S_{\varepsilon_i}\left(\frac{\log(t) - \mu - \boldsymbol{\alpha}^T \boldsymbol{x_i}}{\sigma}\right) = 1 - \Phi\left(\frac{\log(t) - \mu - \boldsymbol{\alpha}^T \boldsymbol{x_i}}{\sigma}\right)$$

Λόγω της έλλειψης κλειστής μορφής της c.h.f. σε αυτή την περίπτωση δεν υφίσταται κάποια απλοποιημένη έκφραση για την εκτίμηση της.

2ο Κεφάλαιο
Μέθοδοι Κρησαρίσματος Μεταβλητών και Ποινικοποίησης σε Περιβάλλοντα Υψηλής Διάστασης

2.1 Εισαγωγή

Στην Ανάλυση Επιβίωσης, έχουν προταθεί διάφορες μέθοδοι επιλογής μεταβλητών με στόχο, στο τελικό στάδιο της διαδικασίας, ο αναλυτής να διαθέτει ένα σύνολο στατιστικά σημαντικών μεταβλητών που επηρεάζουν το χρόνο μελέτης *T*. Στη συνέχεια, το επιλεγμένο αυτό σύνολο μπορεί να χρησιμοποιηθεί ως βάση για την ανάπτυξη κατάλληλου μοντέλου πρόβλεψης. Σύμφωνα με τους Hosmer & Lemeshow (1999), ευρέως διαδεδομένες μέθοδοι επιλογής μεταβλητών – γνωστές και από το πλαίσιο της Κλασικής Γραμμικής Παλινδρόμησης – είναι οι ακόλουθες:

- forward selection: Η διαδικασία ξεκινά από το μηδενικό (null) μοντέλο, το οποίο περιλαμβάνει μόνο το σταθερό όρο. Σε κάθε επόμενο βήμα προστίθεται η επεξηγηματική μεταβλητή που εμφανίζει τη μεγαλύτερη στατιστική σημαντικότητα μεταξύ εκείνων που δεν έχουν ακόμη εισαχθεί στο μοντέλο.
- backward elimination: Η μέθοδος αυτή ακολουθεί την αντίστροφη λογική. Η διαδικασία εκκινεί από το πλήρες (full) μοντέλο, το οποίο περιλαμβάνει όλες τις διαθέσιμες επεξηγηματικές μεταβλητές, καθώς και το σταθερό όρο. Σε κάθε βήμα εξετάζεται η στατιστική σημαντικότητα των συμμεταβλητών και απομακρύνεται εκείνη που δεν πληροί το προκαθορισμένο κριτήριο σημαντικότητας.
- stepwise selection: Αποτελεί συνδυασμό των δύο προηγούμενων προσεγγίσεων. Η μέθοδος επιτρέπει τόσο την προσθήκη όσο και την αφαίρεση μεταβλητών κατά τη διάρκεια της διαδικασίας, λαμβάνοντας υπόψη τη μεταβολή της στατιστικής τους σημαντικότητας μετά την ένταξη νέων χαρακτηριστικών στο μοντέλο.

Οι παραπάνω μέθοδοι εφαρμόζονται κυρίως σε προβλήματα χαμηλής διάστασης, όπου ο αριθμός των επεξηγηματικών μεταβλητών δεν υπερβαίνει το μέγεθος του δείγματος. Αντιθέτως, στο πλαίσιο υψηλής διάστασης δεδομένων, οι αλγόριθμοι αυτοί παρουσιάζουν σημαντικούς περιορισμούς, καθώς δυσκολεύονται να διασφαλίσουν αξιόπιστη και στατιστικά έγκυρη συμπερασματολογία.

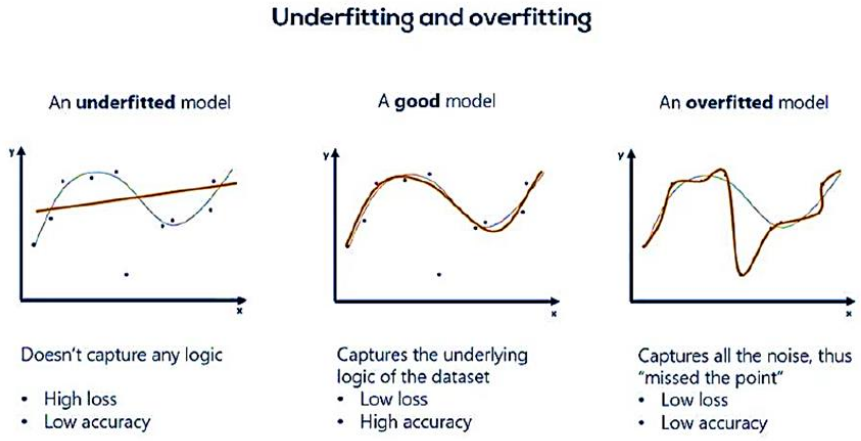
2.2 Υψηλή Διαστατικότητα στα Δεδομένα Επιβίωσης

Στη σύγχρονη εποχή, ένα από τα πιο σημαντικά ζητήματα στο χώρο της Στατιστικής είναι η «εποχή των Μεγάλων Δεδομένων (Big Data)» όπου το πλήθος των παρατηρήσεων (που θα συμβολίζεται με το γράμμα n) σε σχέση με τον αριθμό των μεταβλητών (που θα συμβολίζεται με το γράμμα p) είναι αρκετά μεγάλος, δηλαδή ισχύει $n \gg p$. Σ' αυτό το πλαίσιο, έχουν αναπτυχθεί διάφοροι αλγόριθμοι και μέθοδοι από το πεδίο της Στατιστικής και της Πληροφορικής, οι οποίοι επιτρέπουν την

επεξεργασία μεγάλων ποσοτήτων δεδομένων σε μικρό χρονικό διάστημα. Ωστόσο, το πρόβλημα δεν περιορίζεται μόνο στον όγκο των δεδομένων, αλλά επεκτείνεται και στη συχνότητα με την οποία νέα δεδομένα καταφθάνουν, γεγονός που προσθέτει ακόμη μεγαλύτερες προκλήσεις στους αναλυτές, καθώς τα δεδομένα έρχονται σε πραγματικό χρόνο, ακόμα και σε δευτερόλεπτα.

Αυτό το πρόβλημα έχει αντιμετωπιστεί σε μεγάλο βαθμό με τη χρήση ισχυρών εργαλείων και υπερυπολογιστών. Ωστόσο, ένα νέο και πιο σύνθετο ζήτημα που έχει αναδειχθεί και συζητείται έντονα από τις αρχές του 21ου αιώνα είναι η περίπτωση των δεδομένων υψηλής διάστασης ($p \gg n$), και αυτό ακριβώς είναι το θέμα που θα αναλυθεί στην παρούσα εργασία, καθώς και τρόποι αντιμετώπισής του. Η στατιστική υψηλής διάστασης αναφέρεται στη στατιστική συμπερασματολογία όταν ο αριθμός των επεξηγηματικών μεταβλητών p, που έχει ο αναλυτής στη διάθεσή του, είναι σημαντικά μεγαλύτερος από το μέγεθος του δείγματος n. Στον ιατρικό κλάδο, για παράδειγμα, τα αποτελέσματα επιβίωσης με πολλές επεξηγηματικές μεταβλητές συλλέγονται συστηματικά. Αυτά τα high – dimensional δεδομένα αμφισβητούν τα παραδοσιακά μοντέλα παλινδρόμησης επιβίωσης που εξετάστηκαν στο 1ο Κεφάλαιο, τα οποία είτε είναι αδύνατο να προσαρμοστούν σωστά είτε έχουν χαμηλή προβλεπτική ικανότητα λόγω του φαινομένου της υπερπροσαρμογής.

Με τον όρο υπερπροσαρμογή (overfitting) νοείται το φαινόμενο κατά το οποίο ένα στατιστικό μοντέλο προσαρμόζεται υπερβολικά στα δεδομένα εκπαίδευσης, συλλαμβάνοντας όχι μόνο τη συστηματική δομή της σχέσης μεταξύ μεταβλητών αλλά και τον τυχαίο θόρυβο. Ως αποτέλεσμα, το μοντέλο επιδεικνύει υψηλή προσαρμογή στο αρχικό δείγμα, αλλά χαμηλή ικανότητα γενίκευσης σε νέα, ανεξάρτητα δεδομένα. Πέραν από την υπερπροσαρμογή, ένα επιπλέον σημαντικό ζήτημα είναι η αυξημένη διακύμανση των εκτιμητών των παραμέτρων του μοντέλου, κάτι που καθιστά το μοντέλο ιδιαίτερα ευαίσθητο σε μικρές μετατοπίσεις των τιμών και σχεδόν ανίκανο να εντοπίσει τις πραγματικά στατιστικά σημαντικές μεταβλητές που επηρεάζουν το χρόνο επιβίωσης T.



Εικόνα 2.1. Απεικόνιση των περιπτώσεων υποπροσαρμογής, καλής προσαρμογής και υπερπροσαρμογής σε ένα μοντέλο παλινδρόμησης

Οι Tibshirani & Witten (2010) επισημαίνουν ότι στη γενική περίπτωση υψηλής διάστασης δεδομένων τα προβλήματα είναι αρκετά. Εάν ο πίνακας συνδιακύμανσης των μεταβλητών δεν είναι πλήρους τάξης (δηλαδή υπάρχουν μεταβλητές οι οποίες έχουν τέλεια γραμμική σχέση με κάποιες άλλες, προκαλώντας έτσι το φαινόμενο της πολυσυγγραμμικότητας) δεν μπορεί να ορισθεί αντίστροφός του και συνεπώς οι συντελεστές παλινδρόμησης ελαχίστων τετραγώνων δεν είναι μοναδικοί. Πόσο μάλλον στην Ανάλυση Επιβίωσης, όπου η εκτίμηση των παραμέτρων γίνεται με βάση τη μεγιστοποίηση της μερικής συνάρτησης πιθανοφάνειας που συζητήθηκε στην ενότητα 1.3 και λαμβάνοντας υπόψη τη λογοκρισία των δεδομένων. Οι Tibshirani & Witten (2010) θεώρησαν ότι απαιτείται κάποια μορφή κανονικοποίησης για να μειωθεί η διαστατικότητα του χώρου των συμμεταβλητών. Ακόμα και αν η κανονικοποίηση πραγματοποιηθεί, έτσι ώστε οι συντελεστές παλινδρόμησης να είναι μοναδικοί, η υπερπροσαρμογή αποτελεί σημαντική ανησυχία - υπάρχει κίνδυνος να προσαρμοστεί όχι μόνο το σήμα, αλλά και ο θόρυβος στα δεδομένα, έτσι ώστε το μοντέλο να μην ταιριάζει καλά σε μια νέα παρατήρηση. Απαιτείται μεγάλη προσοχή για να αποφευχθεί η υπερπροσαρμογή, η οποία μπορεί να συμβεί ακόμη και αν $p < n$. Επιπλέον τονίζουν ότι κατά την κατασκευή ενός προγνωστικού μοντέλου που θα είναι χρήσιμο στην αξιολόγηση μελλοντικών ασθενών, μια σημαντική παράμετρος είναι η απλότητα του μοντέλου. Με όλα τα άλλα να είναι ίδια, θα ήταν προτιμότερο ένα μοντέλο που χρησιμοποιεί μόνο ένα μικρό υποσύνολο των επεξηγηματικών μεταβλητών, παρά ένα μοντέλο που τις χρησιμοποιεί όλες. Αυτό συμβαίνει για διάφορους λόγους, όπως ότι θα προσφέρει ευκολία και οικονομία στο γιατρό της μελέτης, καθώς θα χρειαστεί να παρατηρεί κάποιες συγκεκριμένες τιμές σε σχέση με όλες αυτές που υπάρχουν εξαρχής.

Στην Ανάλυση Επιβίωσης, και ειδικότερα στον ιατρικό τομέα, οι επιστήμονες συχνά έρχονται αντιμέτωποι με μεγάλο αριθμό μεταβλητών, καθώς αναλύουν δεδομένα γονιδίων (π.χ. σε πολλά σύνολα δεδομένων υπάρχουν δεκάδες χιλιάδες γονίδια για κάθε ασθενή, ενώ το πλήθος των ατόμων συνήθως είναι μικρότερο, ειδικά στις αρχικές φάσεις των κλινικών δοκιμών). Επίσης, δεδομένα βιοϊατρικής ή ψηφιακής υγείας (που προκύπτουν από φορητές συσκευές υγείας) γίνονται συχνά ultra – high dimensional, καθώς τόσο ο αριθμός των παρατηρήσεων όσο και των μεταβλητών p είναι αρκετά μεγάλος, χωρίς όμως να χάνεται η σχέση $p > n$.

Στο πλαίσιο αυτό, μια αποτελεσματική στρατηγική που έχει προταθεί στη σύγχρονη βιβλιογραφία είναι η εφαρμογή μιας διαδικασίας δύο σταδίων. Αρχικά εφαρμόζεται μια μέθοδος μείωσης διάστασης (feature screening), η οποία περιορίζει τον αριθμό των υποψήφιων μεταβλητών σε ένα διαχειρίσιμο σύνολο. Στη συνέχεια, στο μειωμένο αυτό σύνολο εφαρμόζονται μέθοδοι κανονικοποίησης (regularization), οι οποίες επιτυγχάνουν ταυτόχρονα εκτίμηση παραμέτρων και επιλογή μεταβλητών. Σε κάποιες περιπτώσεις, υπάρχει το ενδεχόμενο τα δύο στάδια να ενοποιούνται σε μία κοινή μεθοδολογική διαδικασία.

2.3 Μέθοδοι Μείωσης Διάστασης Δεδομένων

Η περίοδος 2009 – 2018 χαρακτηρίζεται από έντονη ερευνητική δραστηριότητα στον τομέα της μείωσης διαστάσεων για δεδομένα επιβίωσης υψηλής διάστασης. Οι μέθοδοι που προτάθηκαν διαφοροποιούνται τόσο ως προς το θεωρητικό τους υπόβαθρο όσο και ως προς τις υπολογιστικές τους απαιτήσεις. Στην παρούσα ενότητα παρουσιάζονται συνοπτικά οι κυριότερες τεχνικές της κατηγορίας αυτής, προκειμένου να διαμορφωθεί το απαραίτητο πλαίσιο για την αναλυτική παρουσίαση των πιο σύγχρονων μεθόδων που ακολουθεί.

Ο πρώτος που μελέτησε τη μείωση διαστάσεων στην Ανάλυση Επιβίωσης για δεδομένα υψηλής διάστασης ήταν ο Tibshirani (2009), προτείνοντας έναν συνδυασμό δύο διακριτών τεχνικών: της επιβολής ποινής Lasso (βλ. ενότητα 2.4) και μεθόδων ταξινόμησης από το χώρο της πολυμεταβλητής ανάλυσης. Στόχος της προσέγγισης αυτής είναι η δημιουργία ομάδων μεταβλητών εντός των οποίων οι επεξηγηματικές μεταβλητές εμφανίζονται ανεξάρτητες, ιδιότητα που επιτυγχάνεται μέσω της διαγώνιας διακριτικής ανάλυσης. Στη συνέχεια, εφαρμόζοντας ποινή Lasso εντός κάθε ομάδας, ο νέος εκτιμητής που προκύπτει συνδέεται με τη μέθοδο του συρρικνωμένου κεντροειδούς που προτάθηκε από τους Tibshirani et al. (2002). Η προτεινόμενη μέθοδος, γνωστή ως CUS (Cox Univariate Shrinkage), για την εκτίμηση των συντελεστών $\boldsymbol{\beta}$, βασίζεται στη μεγιστοποίηση της ποσότητας:

$$J(\beta) = \sum_{j=1}^{p} g_j(\beta_j) - \lambda \sum_{j=1}^{p} |\beta_j|, \qquad \text{όπου}$$

$$g_j(\beta_j) = \sum_{k=1}^{K} x_{kj}\beta_j - log \sum_{m \in R_i} e^{x_{mj}\beta_j}$$

Ο δείκτης k εκφράζει την $k-$ οστή ομάδα που προκύπτει από τη διαδικασία ταξινόμησης των p – επεξηγηματικών μεταβλητών σε K ομάδες. Η μεγιστοποίηση της παραπάνω συνάρτησης μπορεί να πραγματοποιηθεί μονοδιάστατα, μέσω της ανεξάρτητης μεγιστοποίησης κάθε όρου της μορφής $g_j(\beta_j) - \lambda|\beta_j|$ ξεχωριστά. Τέλος, αξίζει να σημειωθεί ότι οι εκτιμητές που προκύπτουν εμφανίζουν μεροληψία προς το μηδέν, ιδιαίτερα για μεγάλες τιμές της υπερπαραμέτρου λ που εφαρμόζεται στην ποινή Lasso.

Μια δεύτερη μέθοδος η οποία αποτελεί μία από τις πλέον διαδεδομένες προσεγγίσεις και λειτούργησε ως σημείο αναφοράς για πολλές μεταγενέστερες τεχνικές είναι αυτή των Fan et al. (2010), γνωστή ως Sure Independent Screening (SIS). Η μέθοδος αυτή αποτελεί model – based διαδικασία, βασισμένη στο ημιπαραμετρικό μοντέλο Cox. Η βασική ιδέα της SIS είναι ο υπολογισμός της οριακής (marginal) χρησιμότητας κάθε μεταβλητής X, εξετάζοντας την ξεχωριστά, χωρίς να λαμβάνονται υπόψη οι υπόλοιπες επεξηγηματικές μεταβλητές. Ως οριακή χρησιμότητα ορίζεται ο εκτιμώμενος συντελεστής της X του μοντέλου Cox που προκύπτει από μεγιστοποίηση της μερικής πιθανοφάνειας, όταν στο μοντέλο περιλαμβάνεται μόνο η συγκεκριμένη μεταβλητή. Οι εκτιμήσεις αυτές ταξινομούνται κατά φθίνουσα σειρά σημαντικότητας και επιλέγονται οι πρώτες d. Στις προσομοιώσεις τους, προτάθηκε ως βασική επιλογή για το d ένα πολλαπλάσιο του

$[n/\log(n)]$. Επιπλέον, εισήγαγαν την επαναληπτική εκδοχή της μεθόδου, γνωστή ως Iterative Sure Independent Screening (ISIS), μέσω της οποίας επιτυγχάνεται προοδευτική βελτίωση του συνόλου των επιλεγμένων μεταβλητών και σύγκλιση προς εκείνες που σχετίζονται ουσιαστικά με το χρόνο επιβίωσης T. Στη σχετική βιβλιογραφία έχει αναπτυχθεί εκτενής συζήτηση σχετικά με την επιλογή του αριθμού d των μεταβλητών που διατηρούνται μετά το screening. Σε ορισμένες περιπτώσεις, για να ικανοποιείται η sure screening property ή – σε συνδυασμό με μεθόδους κανονικοποίησης (regularization) – η oracle property, απαιτείται συγκεκριμένος ρυθμός αύξησης του d ως προς το n. Σε απλούστερα θεωρητικά πλαίσια προτείνεται, για μεγάλο n, η επιλογή $d = [n/\log(n)]$, ενώ σε άλλες περιπτώσεις εξετάζεται ακόμη και η επιλογή $d = n - 1$.

Στη λογική της SIS βασίστηκαν οι Zhao et al. (2012), προτείνοντας τη μέθοδο Principled SIS. Πρόκειται επίσης για model – based προσέγγιση στο πλαίσιο του μοντέλου Cox, με τη διαφορά ότι ενσωματώνει την πληροφορία Fisher στη διαδικασία αξιολόγησης των μεταβλητών. Οι Zhao et al. (2012) εισήγαγαν κατώφλι επιλογής της μορφής

$$\gamma_n = c_2 n^{-k}, \qquad k < 1/2\,, c_2 > 0$$

το οποίο βασίζεται στον έλεγχο του αναμενόμενου ποσοστού ψευδώς θετικών μεταβλητών. Με την επιλογή αυτή διασφαλίζονται οι απαιτούμενες θεωρητικές ιδιότητες για αποτελεσματική μείωση διαστάσεων.

Το 2014, η ίδια ομάδα πρότεινε μια νέα, model – free μέθοδο διαλογής, Score Test Screening (STS) (2014), η οποία δεν προϋποθέτει ορισμό συγκεκριμένου μοντέλου, όπως είναι το ημιπαραμετρικό μοντέλο Cox. Στην προσέγγιση αυτή απαιτείται προηγούμενη τυποποίηση των δεδομένων ώστε κάθε μεταβλητή να έχει μέση τιμή 0 και διακύμανση 1. Υπολογίζεται και εδώ μια οριακή στατιστική score για κάθε μεταβλητή ξεχωριστά και επιλέγονται όσες υπερβαίνουν το κατώφλι, το οποίο μπορεί να εκτιμηθεί και με διαδικασίες bootstrap,

$$\gamma_n = c_1 n^{-k}/2, \qquad 0 < \kappa < 1/2.$$

Παράλληλα, προτείνεται επαναληπτική διαδικασία που οδηγεί σε σύγκλιση ενός σταθερού συνόλου σημαντικών μεταβλητών. Η μέθοδος αυτή διαμορφώνει ένα γενικό πλαίσιο διαλογής βασισμένο σε score test, το οποίο είναι ευρέως εφαρμόσιμο, υπολογιστικά αποδοτικό και θεωρητικά συνεπές ως προς τη sure screening property. Επιπλέον, εισάγεται διαδικασία αναδειγματοληψίας για την επιλογή του τελικού αριθμού μεταβλητών μέσω της έννοιας της αναπαραγώγιμης διαλογής (reproducible screening). Τέλος, προτείνεται μια επαναληπτική μέθοδος διαλογής βασισμένη σε projected subgradient methods από τη βιβλιογραφία αριθμητικής βελτιστοποίησης, η οποία συνδέεται στενά με τεχνικές αραιής παλινδρόμησης.

Την ίδια χρονιά οι Song et al. (2014) πρότειναν μία model – free μέθοδο λογοκριμένης ανεξαρτησίας κατάταξης (censored rank independence), βασισμένη στο δείκτη τ_κ. Ο δείκτης αυτός χρησιμοποιεί την εκτίμηση Kaplan – Meier για τη συνάρτηση επιβίωσης s.f. και μπορεί να θεωρηθεί ως ένας αντίστροφος δείκτης του συντελεστή συσχέτισης Kendall τ, προσαρμοσμένος για λογοκριμένα δεδομένα. Η επιλογή των μεταβλητών πραγματοποιείται με κατώφλι όμοιο με εκείνο της μεθόδου Principled SIS, ενώ η χρήση του συντελεστή Kendall επιτρέπει τον εντοπισμό

μονοτονικών μετασχηματισμών μεταξύ αποκρίσεων και επεξηγηματικών μεταβλητών, καθώς και πιθανών μη γραμμικών σχέσεων. Επιπλέον, η μέθοδος αποδεικνύεται ανθεκτική σε ακραίες τιμές λόγω της ιδιότητας του συντελεστή Kendall και διατηρεί τη sure screening property χωρίς να απαιτούνται αυστηρές υποθέσεις σχετικά με τις ουρές των κατανομών.

Το 2016 υπήρξε μια χρονιά κατά την οποία προτάθηκαν αρκετές νέες μέθοδοι feature screening, οι οποίες διαφοροποιούνται ως προς τη μεθοδολογική τους προσέγγιση. Αρχικά, οι Li et al. (2016) πρότειναν μία νέα model – free μέθοδο, βασισμένη στο survival impact index (ξ). Πρόκειται για την πρώτη μέθοδο με αναφορά σε ultra – high dimensional δεδομένα επιβίωσης, δηλαδή περιβάλλοντα όπου ο αριθμός των μεταβλητών ξεπερνά το μέγεθος του δείγματος, το οποίο ωστόσο είναι μεγαλύτερο από αυτό που είχε χρησιμοποιηθεί σε προηγούμενες προσεγγίσεις. Ο survival impact index ερμηνεύεται ως η σταθμισμένη μέση απόλυτη διαφορά μεταξύ της εκτιμώμενης συνάρτησης επιβίωσης Kaplan – Meier με και χωρίς τη συγκεκριμένη επεξηγηματική μεταβλητή X. Ο δείκτης λαμβάνει τιμές μεταξύ 0 και 1 όπως μία πιθανότητα, γεγονός που διευκολύνει την ερμηνεία και μειώνει τη μεταβλητότητα κατά τη διαδικασία διαλογής. Επιπλέον, οι υπολογισμοί γίνονται σε συγκεκριμένο εύρος χρόνου αποτυχίας και η μέθοδος διατηρεί τη sure screening property.

Στη συνέχεια, οι Yang et al. (2016) πρότειναν τη μέθοδο Sure Joint Screening (SJS) για ultra – high dimensional δεδομένα επιβίωσης. Η μέθοδος τροποποιεί τη μερική πιθανοφάνεια έτσι ώστε να είναι δυνατή η εκτίμηση των συντελεστών όταν ο αριθμός των πραγματικά σημαντικών μεταβλητών είναι μικρότερος από μια σταθερά m, αντίστοιχη με την τιμή $d = [n/\log(n)]$ που ορίσθηκε και χρησιμοποιείται στις SIS και ISIS μεθόδους. Σε περιβάλλον χαμηλής διάστασης, η εκτίμηση των β μέσω της μερικής πιθανοφάνειας δεν παρουσιάζει προβλήματα. Σε περιβάλλον υψηλής διάστασης, όμως, η εκτίμηση της δεύτερης παραγώγου καθίσταται ανέφικτη, καθιστώντας τη μεγιστοποίηση προβληματική. Για την αντιμετώπιση του ζητήματος αυτού, οι Yang et al. (2016) χρησιμοποιούν ανάπτυγμα Taylor και δημιουργούν έναν επαναληπτικό αλγόριθμο που στοχεύει στη σύγκλιση προς τον πραγματικό αριθμό μη μηδενικών συντελεστών. Η μέθοδος αξιοποιεί όλη την πληροφορία ταυτόχρονα (joint) αντί να εξετάζει κάθε μεταβλητή ξεχωριστά (marginal), χρησιμοποιώντας την τροποποιημένη μερική πιθανοφάνεια ως joint score για την επιλογή των σημαντικών μεταβλητών. Με αυτόν τον τρόπο, η SJS διατηρεί τη sure screening property και εντοπίζει αποτελεσματικά τους πραγματικά σημαντικούς συντελεστές.

Η τρίτη μέθοδος, που προτάθηκε την ίδια χρονιά, προέρχεται από τους Hong et al. (2016). Πρόκειται για μία υπό συνθήκη model – based μέθοδος διαλογής γνωστή ως Cox Conditional Screening method (CoxCS) η οποία βασίζεται στο ημιπαραμετρικό μοντέλο Cox και απευθύνεται σε δεδομένα εξαιρετικά υψηλής διάστασης. Η προσέγγιση αυτή βασίζεται στην ιδέα των Barut et al. (2012) οι οποίοι εφάρμοσαν μία υπό συνθήκη εκδοχή της μεθόδου SIS σε μοντέλα GLM (Generalized Linear Model). Τα μοντέλα GLM αποτελούν γενίκευση των μοντέλων παλινδρόμησης, στα οποία η μεταβλητή απόκρισης μπορεί να μην ακολουθεί τις υποθέσεις της κλασικής γραμμικής παλινδρόμησης και ακολουθεί κατανομή που ανήκει στην εκθετική

οικογένεια κατανομών. Η μέθοδος CoxCS αξιοποιεί εκ των προτέρων πληροφορία σε ό,τι ορισμένες μεταβλητές σχετίζονται με το χρόνο επιβίωσης. Αποτελεί ουσιαστικά μία από τις πρώτες προσεγγίσεις που εισάγουν εκ των προτέρων γνώση (prior knowledge) στη διαδικασία διαλογής. Οι μεταβλητές αυτές εντάσσονται σε ένα υποσύνολο C. Στη συνέχεια, προτείνουν την εφαρμογή μοντέλων Cox που περιλαμβάνουν τις μεταβλητές του υποσυνόλου C και, κάθε φορά, μια μεταβλητή από το συμπληρωματικό υποσύνολο C΄, το οποίο περιλαμβάνει τις μεταβλητές για τις οποίες δεν υπάρχει αρχικά ένδειξη συσχέτισης με το χρόνο επιβίωσης. Από τα μοντέλα αυτά λαμβάνονται οι εκτιμήσεις των συντελεστών των μεταβλητών του υποσυνόλου C΄ και επιλέγονται όσες υπερβαίνουν το κατώφλι γ. Το κατώφλι που προτείνεται είναι αντίστοιχο με εκείνο που χρησιμοποιείται στις μεθόδους SIS και ISIS. Η συγκεκριμένη μέθοδος παρουσιάζει ορισμένα βασικά πλεονεκτήματα σε σχέση με προηγούμενες προσεγγίσεις. Είναι υπολογιστικά αποδοτική και μπορεί να εντοπίσει σημαντικές μεταβλητές ακόμη και όταν το σήμα επίδρασης τους στο χρόνο επιβίωσης δεν είναι ισχυρό μεμονωμένα, αλλά προκύπτει μέσα από τη συνδυασμένη τους επίδραση (jointly important signals). Επιπλέον, διατηρεί την ιδιότητα sure screening property, ενώ οι δύο εναλλακτικές προσεγγίσεις που προτείνονται από τους συγγραφείς οδηγούν σε πολύ μικρό ποσοστό ψευδώς επιλεγμένων μεταβλητών. Αξίζει να σημειωθεί ότι στην περίπτωση όπου το υποσύνολο $C = \emptyset$, η μέθοδος είναι ουσιαστικά ισοδύναμη με τη μέθοδο SIS των Fan et al. (2010).

Τον επόμενο χρόνο, οι Ma et al. (2017) παρουσίασαν μια συνολική διαδικασία επιλογής μεταβλητών (variable selection) για ultra – high dimensional δεδομένα επιβίωσης, η οποία χωρίζεται σε δύο σκέλη: feature screening και penalization. Στόχος είναι η δημιουργία ενός μοντέλου πρόβλεψης του χρόνου επιβίωσης T με τις πραγματικά σημαντικές μεταβλητές. Η μέθοδος βασίζεται στον C – Statistic (concordance measure) που μετρά τη συμφωνία κατάταξης μεταξύ των επεξηγηματικών μεταβλητών και του χρόνου επιβίωσης – ένα τυπικό βήμα για την αξιολόγηση της προγνωστικής ακρίβειας ενός μοντέλου επιβίωσης. Εισάγουν ένα smoothed C – Statistic, κατάλληλο για ultra – high dimensional δεδομένα, το οποίο ενσωματώνει συνολικά την πληροφορία των δεδομένων. Η μεγιστοποίηση αυτού του μέτρου οδηγεί στην εκτίμηση των συντελεστών των επεξηγηματικών μεταβλητών, δεδομένου ότι υπάρχει ένας αριθμός μεταβλητών με μη μηδενικούς συντελεστές. Η μέθοδος διατηρεί τη sure screening property για συγκεκριμένο κατώφλι, ανάλογο με αυτό που χρησιμοποιήθηκε στις SIS και ISIS μεθόδους. Επιπλέον, οι Ma et al. (2017) προτείνουν έναν επαναληπτικό αλγόριθμο, που οδηγεί στη σύγκλιση των πραγματικά σημαντικών μεταβλητών. Συνδυάζοντας το smoothed C – statistic με τη SCAD ποινή (βλ. ενότητα 2.4) προκύπτει ένα αραιό μοντέλο παλινδρόμησης, το οποίο περιλαμβάνει μόνο τις στατιστικά σημαντικές μεταβλητές που επηρεάζουν το χρόνο επιβίωσης. Επιπλέον, αποδεικνύεται ότι η μέθοδος ικανοποιεί και την oracle property για την ποινή παλινδρόμησης.

Τέλος, το 2018 προτάθηκαν δύο νέες model – free μέθοδοι feature screening, βασισμένες αντίστοιχα στη στατιστική συνάρτηση Kolmogorov του μη παραμετρικού ελέγχου καλής προσαρμογής και στο συντελεστή συσχέτισης Pearson. Οι Hong et al. (2018) πρότειναν μια διαφορετική προσέγγιση βασισμένη στη συνάρτηση του

Kolmogorov, υπολογίζοντας τη διαφορά μεταξύ των συναρτήσεων πυκνότητας – πυρήνα του χρόνου T (kernel density function of T), υψωμένων σε μια δύναμη $\gamma > 0$ (στις προσομοιώσεις της εργασίας προτείνονται τιμές κοντά στη μονάδα), για κάθε επεξηγηματική μεταβλητή σε διαφορετικά στρώματα ή επίπεδα. Ο δείκτης που προκύπτει από αυτή τη διαδικασία ονομάζεται IPOD (Integrated Power Density). Στη συγκεκριμένη διαδικασία είναι απαραίτητη η κατηγοριοποίηση των συνεχών μεταβλητών, ώστε ο δείκτης να είναι υπολογιστικά διαχειρίσιμος. Αφού υπολογισθεί η τιμή του δείκτη για κάθε επεξηγηματική μεταβλητή, αυτή συγκρίνεται με το κατώφλι cn^{-v} όπου $c > 0$ και $0 < v < 1/2$. Μεταβλητές των οποίων η τιμή υπερβαίνει το κατώφλι, θεωρούνται σημαντικές για την ερμηνεία του χρόνου επιβίωσης. Οι Hong et al. (2018) προτείνουν επίσης συγκεκριμένους τρόπους κατηγοριοποίησης των συνεχών μεταβλητών, όπως η κατηγοριοποίηση βάσει ποσοστημορίων. Ο αριθμός των ποσοστημορίων που χρησιμοποιούνται σε κάθε μεταβλητή καθορίζεται μέσω μιας διαδικασίας που προτείνεται στην εργασία τους. Μεταξύ των βασικών πλεονεκτημάτων της μεθόδου συγκαταλέγονται η ικανοποίηση της sure screening property, η μη παραμετρική φύση της προσέγγισης, καθώς και η δυνατότητα χειρισμού διαφορετικών τύπων επεξηγηματικών μεταβλητών (συνεχών και διακριτών). Τα χαρακτηριστικά αυτά προσδίδουν μεγαλύτερη ευελιξία και αυξημένη ικανότητα εντοπισμού των πραγματικά σημαντικών μεταβλητών σε περιβάλλοντα εξαιρετικά υψηλής διάστασης.

Οι Hao et al. (2018) πρότειναν έναν δείκτη βασισμένο στο συντελεστή συσχέτισης Pearson. Συγκεκριμένα, για κάθε επεξηγηματική μεταβλητή υπολογίζεται ο συντελεστής Pearson μεταξύ της μεταβλητής και του χρόνου επιβίωσης T. Η εκτίμηση γίνεται μέσω της εμπειρικής συνάρτησης κατανομής της επεξηγηματικής μεταβλητής και της εκτίμησης Kaplan – Meier της συνάρτησης επιβίωσης του χρόνου T. Το κατώφλι επιλογής ορίζεται με τον ίδιο τρόπο όπως στις μεθόδους SIS και ISIS, και κάθε μεταβλητή της οποίας η απόλυτη τιμή του δείκτη υπερβαίνει το κατώφλι περιλαμβάνεται στο σύνολο M των σημαντικών μεταβλητών για το χρόνο επιβίωσης. Η μέθοδος ικανοποιεί τη sure screening property και διαχειρίζεται αποτελεσματικά το ποσοστό των ψευδώς επιλεγμένων μεταβλητών. Επίσης, οι Hao et al. (2018) απέδειξαν ότι η μέθοδος είναι συνεπής σε κατάταξη χωρίς να απαιτούνται υποθέσεις πεπερασμένης ροπής για τις προγνωστικές μεταβλητές ή την απόκριση. Λαμβάνοντας υπόψη ότι κάποιες μεταβλητές μπορεί να είναι οριακά μη σημαντικές όταν εξετάζονται ξεχωριστά, αλλά σημαντικές σε συνδυασμό με άλλες, προτάθηκε μια επαναληπτική διαδικασία της μεθόδου, ώστε να εντοπίζονται και τέτοιες περιπτώσεις και να μειώνεται η πιθανότητα απώλειας σημαντικών μεταβλητών.

Οι μέθοδοι που παρουσιάστηκαν μέχρι στιγμής παρέχουν ένα στέρεο θεωρητικό και υπολογιστικό πλαίσιο για τη μείωση διαστάσεων σε δεδομένα επιβίωσης υψηλής διάστασης. Παράλληλα, η ποικιλία των προσεγγίσεων, από model – based έως model – free τεχνικές, υπογραμμίζει την εξέλιξη της βιβλιογραφίας και τις συνεχείς προσπάθειες βελτίωσης, τόσο της ακρίβειας επιλογής σημαντικών μεταβλητών, όσο και της υπολογιστικής αποτελεσματικότητας των μεθόδων. Στη συνέχεια, θα παρουσιασθούν πιο σύγχρονες και λεπτομερώς αναλυμένες τεχνικές, από το 2020 και μέχρι σήμερα, οι οποίες αποτελούν βασικό άξονα της παρούσας εργασίας.

2.3.1 Μέθοδος διαλογής μεταβλητών με νόρμα L_q

Οι Hong et al. (2020) πρότειναν μια υβριδική διαδικασία διαλογής μεταβλητών βασισμένη στον κανόνα L_q, η οποία παρουσιάζει ομοιότητες με την προσέγγιση των Hong et al. (2018). Ορίζουν το γενικό L_q κανόνα μιας συνάρτησης $g(T)$ ως εξής:

$$\left|\left|g(T)\right|\right|_q = \{E(|g(T)|^q)\}^{1/q} = \left\{-\int_0^\infty |g(t)|^q dS(t)\right\}^{1/q}$$

όπου $q \geq 1$ και η τελευταία ισότητα προκύπτει από τη σχέση $-dS(t) = f(t)dt$. Όπως και στην εργασία των Hong et al. (2018), η μέθοδος εφαρμόζεται αρχικά σε κατηγορικές επεξηγηματικές μεταβλητές με, έστω, K επίπεδα. Για κάθε μεταβλητή X_j και για όλους τους συνδυασμούς επιπέδων $\kappa_1 \neq \kappa_2$, υπολογίζεται η ποσότητα:

$$\Psi_j^{(q)} = \max_{k_1,k_2 \in \{1,\dots,K_j\}} \left|\left|S(T|X_j = k_1) - S(T|X_j = k_2)\right|\right|_q$$

Η ποσότητα αυτή μετρά τη μέγιστη απόσταση μεταξύ των συναρτήσεων επιβίωσης που αντιστοιχούν σε διαφορετικά επίπεδα της μεταβλητής. Στην περίπτωση όπου $\Psi_j^{(q)} = 0$ αυτό αποτελεί ένδειξη ότι ο χρόνος επιβίωσης είναι ανεξάρτητος από τη μεταβλητή X_j. Στην πράξη, όταν στα δεδομένα υπάρχουν d παρατηρούμενοι χρόνοι αποτυχίας, η παραπάνω ποσότητα εκτιμάται ως:

$$\widehat{\Psi}_j^{(q)} = \max_{k_1,k_2 \in \{1,\dots,K_j\}} \left\{-\int_0^\infty \left|\hat{S}(t|X_j = k_1) - \hat{S}(t|X_j = k_2)\right|^q d\hat{S}(t)\right\}^{1/q}$$

$$= \max_{k_1,k_2 \in \{1,\dots,K_j\}} \left[\sum_{l=1}^{d} \left|\hat{S}(t_l|X_j = k_1) - \hat{S}(t_l|X_j = k_2)\right|^q \{\hat{S}(t_{l-1}) - \hat{S}(t_l)\}\right]^{1/q}$$

όπου $\hat{S}(\cdot)$ ο εκτιμητής Kaplan – Meier και $t_0 = 0$ για λόγους ευκολίας. Έστω M το σύνολο των πραγματικά ενεργών επεξηγηματικών μεταβλητών που σχετίζονται με το χρόνο επιβίωσης. Με βάση την εκτιμώμενη ποσότητα $\widehat{\Psi}_j^{(q)}$, το εκτιμώμενο σύνολο ενεργών μεταβλητών ορίζεται ως:

$$\widehat{M} = \left\{j: \widehat{\Psi}_j^{(q)} > cn^{-v}, j = 1, \dots, p\right\}$$

όπου $q \geq 1$ είναι ο κανόνας που επιλέγεται από τον ερευνητή, ενώ $c > 0 \; και \; v \in [0,1/2)$ είναι προκαθορισμένες σταθερές τέτοιες ώστε $\min_{j \in M} \Psi_j^{(q)} \geq 2cn^{-v}$. Στην πράξη, οι επεξηγηματικές μεταβλητές μπορεί να είναι όχι μόνο κατηγορικές αλλά και συνεχείς. Για την αντιμετώπιση αυτής της περίπτωσης, προτείνουν μια διαδικασία κατηγοριοποίησης κάθε συνεχούς μεταβλητής X_j σε K_j επίπεδα. Η νέα τιμή, που συμβολίζεται ως $\tilde{X}_j$, λαμβάνει την τιμή k αν και μόνο αν:

$$X_j \in [\hat{Q}_{j(k-1)}, \hat{Q}_{j(k)})$$

όπου $\hat{Q}_{j(k)}$ είναι το $100 \times \kappa/K_j$ ποσοστημόριο της εμπειρικής κατανομής της X_j. Για λόγους ευκολίας, τίθεται $\hat{Q}_{j(0)} = -\infty$ και $\hat{Q}_{j(K_j)} = +\infty$. Η κατηγοριοποίηση αυτή

μπορεί να πραγματοποιηθεί με διαφορετικούς τρόπους. Αν υπάρχουν N διαφορετικά σχήματα διακριτοποίησης, η ποσότητα $\Psi_j^{(q)}$ επαναδιατυπώνεται ως:

$$\widetilde{\Psi}_j^{(q)} = \sum_{u=1}^{N} \widehat{\Psi}_{j,\Lambda_{ju}}^{(q)}$$

όπου

- $\widehat{\Psi}_{j,\Lambda_{ju}}^{(q)} = \max_{k_1,k_2\in\{1,\dots,K_j\}} \left[\sum_{l=1}^{d} \left|\hat{S}\left(t_l \middle| \tilde{X}_{ju} = k_1\right) - \hat{S}\left(t_l \middle| \tilde{X}_{ju} = k_2\right)\right|^q \left\{\hat{S}(t_{l-1}) - \hat{S}(t_l)\right\}\right]^{\frac{1}{q}}$
- $\Lambda_{ju} = \left\{\left[\widehat{Q}_{ju(k-1)}, \widehat{Q}_{ju(k)}\right) : k = 1, \dots, K_{ju} \text{ και } \bigcup_k^{K_{ju}} \left[\widehat{Q}_{ju(k-1)}, \widehat{Q}_{ju(k)}\right) = \mathbb{R}\right\}$: διαφορετικό σχήμα διακριτοποίησης της μεταβλητής.

Για να εξασφαλιστεί ότι κάθε διακριτοποίηση περιλαμβάνει επαρκή αριθμό διαστημάτων, ορίζεται η κάτωθι ποσότητα

$$K_{ju} = 3, \dots, [log(n)]$$

οπότε προκύπτουν $N = [log(n) - 2]$ διαφορετικά σχήματα διακριτοποίησης. Μια μεταβλητή θεωρείται ανεξάρτητη από το χρόνο επιβίωσης μόνο όταν για όλα τα σχήματα διακριτοποίησης ισχύει $\widehat{\Psi}_{j,\Lambda_{ju}}^{(q)} = 0$. Με βάση τον παραπάνω ορισμό, το τελικό σύνολο ενεργών μεταβλητών ορίζεται ως:

$$\widetilde{M} = \left\{j : \widetilde{\Psi}_j^{(q)} > \tilde{c}n^{-\tilde{\nu}}, j = 1, \dots, p\right\}$$

όπου $\tilde{c} > 0$ και $\tilde{\nu} \in [0,1/2)$ είναι κατάλληλες σταθερές τέτοιες ώστε $\min_{j\in M} \Psi_{j0}^{(q)} \geq 2\tilde{c}n^{-\tilde{\nu}}$. Ένα σημαντικό ζήτημα της μεθόδου είναι η επιλογή της τιμής του q. Για να μην εξαρτάται η διαδικασία από μια συγκεκριμένη επιλογή, οι Hong et al. (2020) προτείνουν έναν υβριδικό μηχανισμό μάθησης βασισμένο σε πολλαπλές τιμές του q. Έστω

$$1 \leq q_1 < \cdots < q_L < \infty$$

ένα σύνολο πιθανών τιμών. Το υβριδικό σύνολο μεταβλητών ορίζεται ως:

$$\widetilde{M}_h = \bigcup_{l=1}^{L} \widetilde{M}^{(ql)} \quad \text{όπου}$$

$$\widetilde{M}^{(ql)} = \left\{j : \widetilde{\Psi}_j^{(ql)} > \tilde{c}_h n^{-v_l}, j = 1, \dots, p\right\}$$

με $v_l \in [0,1/2)$: σταθερά που εξαρτάται από το q_l και $\tilde{c}_h > 0$ σταθερά ανεξάρτητη από το l. Οι Hong et al. (2020) προτείνουν οι τιμές q_l να προκύπτουν από την ακολουθία Fibonacci, έτσι ώστε για ακραίες τιμές του q να προσεγγίζονται γνωστές στατιστικές αποστάσεις, όπως η στατιστική Cramér – von Mises ($q = 2$) και η στατιστική Kolmogorov ($q = \infty$), Μέσω προσομοιώσεων έδειξαν ότι για τιμές $q > 30$ η προκύπτουσα στατιστική προσεγγίζει πολύ καλά τη στατιστική Kolmogorov. Για τον λόγο αυτό προτείνεται η χρήση τιμών της ακολουθίας Fibonacci έως το 89. Αναλυτικές πληροφορίες σχετικά με την επιλογή των σταθερών c, v, $\tilde{c}$, $\tilde{\nu}$, $\tilde{c}_h$ και v_l παρουσιάζονται στα επόμενα μέρη της εργασίας τους, όπου διατυπώνονται οι

απαραίτητες συνθήκες, καθώς και σχετικά θεωρήματα και λήμματα που τεκμηριώνουν τις θεωρητικές ιδιότητες της μεθόδου.

Μεταξύ των βασικών πλεονεκτημάτων της προσέγγισης συγκαταλέγονται η ικανοποίηση της sure screening property, η απουσία παραμετρικών υποθέσεων και η αμεταβλητότητα της μεθόδου ως προς μονοτονικούς μονομεταβλητούς μετασχηματισμούς τόσο στο χρόνο επιβίωσης όσο και στις επεξηγηματικές μεταβλητές. Επιπλέον, η υβριδική χρήση πολλαπλών τιμών του q συμβάλλει στη μείωση των ψευδώς αρνητικών αποτελεσμάτων, σε αντίθεση με ορισμένες εργασίες που στοχεύουν κυρίως στη μείωση των ψευδώς θετικών αποτελεσμάτων. Τέλος, καθώς η προτεινόμενη διαδικασία βασίζεται σε εκτιμητές Kaplan – Meier, ο υπολογισμός της είναι σχετικά απλός και μπορεί να εφαρμοστεί αποτελεσματικά σε δεδομένα εξαιρετικά υψηλής διάστασης.

2.3.2 Μέθοδος Κρησαρίσματος με χρήση Λογιστικής Παλινδρόμησης σε ζευγοποιημένες παρατηρήσεις ασθενών - μαρτύρων

Οι Yi et al. (2021) προτείνουν μια μεθοδολογική προσέγγιση η οποία συνδυάζει διαφορετικά στατιστικά εργαλεία που χρησιμοποιούνται ευρέως στην ανάλυση δεδομένων επιβίωσης, ιδιαίτερα σε εφαρμογές του ιατρικού χώρου. Συγκεκριμένα, η προτεινόμενη μεθοδολογία βασίζεται στο συνδυασμό του αναλογικού μοντέλου κινδύνου του Cox, του nested case – control Cox μοντέλου παλινδρόμησης (NCC), καθώς και της λογιστικής παλινδρόμησης matched case – control. Στα time – to – event δεδομένα, ο διαχωρισμός των παρατηρήσεων σε ομάδες ασθενών και μαρτύρων δεν αποτελεί τη συνήθη πρακτική, καθώς η βασική διάκριση αφορά κυρίως στο εάν μια παρατήρηση είναι λογοκριμένη ή όχι. Ωστόσο, οι Yi et al. (2021) αξιοποιούν στοιχεία από τις παραπάνω μεθοδολογίες ώστε να διαμορφώσουν μια διαδικασία μείωσης διάστασης (feature screening) κατάλληλη για περιπτώσεις όπου το μέγεθος του δείγματος n είναι μεγάλο και ο αριθμός των επεξηγηματικών μεταβλητών p αυξάνεται με ρυθμό μη πολυωνυμικής τάξης ως προς το n (NP – διάστασης). Η προτεινόμενη μέθοδος είναι ιδιαίτερα κατάλληλη για τη μελέτη σπάνιων ασθενειών, όπου το ποσοστό των γεγονότων αποτυχίας είναι σχετικά μικρό. Στη συνέχεια παρουσιάζονται συνοπτικά οι βασικές τεχνικές στις οποίες βασίζεται η προτεινόμενη προσέγγιση, ώστε να καταστεί σαφής η διαδικασία κατασκευής της τελικής μεθοδολογίας.

2.3.2.1 Αναλογικό Μοντέλο Κινδύνου του Cox με βάση την εργασία των Yi et al.

Έστω διαθέσιμα δεδομένα με n παρατηρήσεις και p επεξηγηματικές μεταβλητές. Οι βασικοί συμβολισμοί που χρησιμοποιούνται είναι οι ακόλουθοι:

- T_i : χρόνος επιβίωσης του $i-οστού$ ατόμου
- C_i: χρόνος λογοκρισίας του $i-οστού$ ατόμου
- $\delta_i = I(T_i \leq C_i)$: δείκτης αποτυχίας (1 εάν η παρατήρηση δεν είναι λογοκριμένη)
- $t_i = \min(T_i, C_i)$: παρατηρούμενος χρόνος
- $Y_i(t) = I(t_i \geq t)$: δείκτης κινδύνου του $i-οστού$ ατόμου τη χρονική στιγμή t.
- $R_i = \{j : Y_j(T_i) = 1\}$: σύνολο κινδύνου λίγο πριν από το χρόνο αποτυχίας T_i.

Διαχωρίζουν το σύνολο των επεξηγηματικών μεταβλητών σε δύο κατηγορίες:

- $\boldsymbol{X}_i$: διάνυσμα συνεχών μεταβλητών διαστάσεων $p_x \times 1$
- $\boldsymbol{Z}_i$: διάνυσμα κατηγορικών μεταβλητών διαστάσεων $p_Z \times 1$.

όπου ισχύει

$$p_x + p_Z = p$$

Η διαφοροποίηση αυτής της διατύπωσης σε σχέση με την κλασική παρουσίαση του μοντέλου Cox στην υποενότητα 1.5.1 έγκειται στο ρητό διαχωρισμό των συνεχών και κατηγορικών μεταβλητών, με αντίστοιχα διανύσματα συντελεστών β_X και β_Z. Το αναλογικό μοντέλο κινδύνου λαμβάνει τη μορφή:

$$\lambda(t|X_i, Z_i) = \lambda_0(t)\exp(\boldsymbol{X}_i^T\boldsymbol{\beta}_x + \boldsymbol{Z}_i^T\boldsymbol{\beta}_z) \tag{2.1}$$

ή ισοδύναμα

$$\lambda(t|X_i, Z_i) = \lambda_0(t)\exp\left\{\sum_{l=1}^{p} W_{i(l)}\beta_l\right\} \tag{2.2}$$

όπου $W_i = (\boldsymbol{X}_i^T, \boldsymbol{Z}_i^T)^T$. Στο πλαίσιο δεδομένων υψηλής διάστασης, οι Yi et al. (2021) προτείνουν μια διαδικασία feature screening κατά την οποία οι μεταβλητές ταξινομούνται βάσει της στατιστικής Wald. Για τον υπολογισμό της στατιστικής αυτής εφαρμόζεται ξεχωριστό μοντέλο Cox για κάθε μεταβλητή. Στη συνέχεια απορρίπτεται το κατώτερο ποσοστό $100(1-\alpha)\%$ των μεταβλητών, όπου $\alpha \in (0,1)$. Η μερική συνάρτηση πιθανοφάνειας του μοντέλου Cox δίνεται από:

$$L_{partial}(\boldsymbol{\beta}) = \prod_{i=1}^{n}\left\{\frac{\exp(X_i^T\boldsymbol{\beta}_X + Z_i^T\boldsymbol{\beta}_Z)}{\sum_{j\in R_i}\exp\left(X_j^T\boldsymbol{\beta}_X + Z_j^T\boldsymbol{\beta}_Z\right)}\right\}^{\delta_i} \tag{2.3}$$

και οι αντίστοιχοι εκτιμητές συμβολίζονται ως $\widehat{\boldsymbol{\beta}}_{partial}$. Ωστόσο, σε προβλήματα πολύ υψηλής διάστασης, ο υπολογισμός της παραπάνω πιθανοφάνειας μπορεί να είναι ιδιαίτερα απαιτητικός υπολογιστικά. Το πρόβλημα αυτό γίνεται εντονότερο σε μελέτες σπάνιων ασθενειών, όπου το μεγαλύτερο μέρος των παρατηρήσεων είναι λογοκριμένο. Στην περίπτωση αυτή, η χρήση όλων των υποκειμένων στο σύνολο κινδύνου συμβάλλει ελάχιστα στη βελτίωση των εκτιμήσεων των παραμέτρων, ενώ αυξάνει σημαντικά το υπολογιστικό κόστος της διαδικασίας. Για τον λόγο αυτό, ενσωματώνουν στη διαδικασία την τεχνική NCC.

2.3.2.2 Μοντέλο παλινδρόμησης Cox σε εμφωλευμένη μελέτη ασθενών – μαρτύρων

Στο πλαίσιο της μεθόδου αυτής, οι παρατηρήσεις διαχωρίζονται σε δύο ομάδες:

- case**:** παρατηρήσεις με πλήρη χρόνο επιβίωσης ($\delta_i = 1$)
- control**:** παρατηρήσεις με λογοκριμένο χρόνο ($\delta_i = 0$), οι οποίες χρησιμοποιούνται ως ομάδα σύγκρισης για τις παρατηρήσεις της ομάδας case.

Έστω ότι υπάρχουν M πλήρεις χρόνοι επιβίωσης χωρίς δεσμούς (βλ. Ενότητα 1.3), οι οποίοι διατάσσονται ως

$$T_1 < T_2 < \cdots < T_M$$

Για λόγους ευκολίας θεωρείται ότι οι πρώτες M παρατηρήσεις του δείγματος αντιστοιχούν στις περιπτώσεις αποτυχίας. Για κάθε άτομο της ομάδας case

επιλέγονται m άτομα από την ομάδα control χωρίς επανάθεση. Η παράμετρος m αποτελεί μια μικρή προκαθορισμένη θετική σταθερά, συνήθως μεταξύ 1 και 5, και η επιλογή της εξαρτάται και από το μέγεθος του δείγματος. Για πλήρη κάλυψη της διαδικασίας, προτείνουν την εφαρμογή διαφόρων τιμών της παραμέτρου m πριν από την ολοκλήρωση της όποιας τελικής απόφασης. Έστω R_i^* το σύνολο των επιλεγμένων control παρατηρήσεων για το $i-οστό$ άτομο της case ομάδας και

$$\tilde{R}_i^* = R_i^* \cup \{i\}$$

Αντικαθιστώντας το σύνολο R_i με το νέο σύνολο $\tilde{R}_i^*$, προκύπτει η NCC μερική πιθανοφάνεια:

$$L_{NCC}(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left\{ \frac{\exp\left(X_i^T \boldsymbol{\beta}_X + Z_i^T \boldsymbol{\beta}_Z\right)}{\sum_{j \in \tilde{R}_i} \exp\left(X_j^T \boldsymbol{\beta}_X + Z_j^T \boldsymbol{\beta}_Z\right)} \right\}^{\delta_i} = \prod_{i=1}^{M} \frac{\exp\left(X_i^T \boldsymbol{\beta}_X + Z_i^T \boldsymbol{\beta}_Z\right)}{\sum_{j \in \tilde{R}_i} \exp\left(X_j^T \boldsymbol{\beta}_X + Z_j^T \boldsymbol{\beta}_Z\right)} \qquad (2.4)$$

Οι αντίστοιχοι εκτιμητές των παραμέτρων συμβολίζονται ως $\widehat{\boldsymbol{\beta}}_{NCC}$. Μια σημαντική παρατήρηση που επισημαίνεται είναι ότι, υπό κατάλληλες συνθήκες κανονικότητας, οι εκτιμήσεις των συντελεστών του μοντέλου Cox μπορούν να προκύψουν είτε από τη μερική πιθανοφάνεια του πλήρους μοντέλου (2.3) είτε από την αντίστοιχη πιθανοφάνεια NCC (2.4), με τις δύο εκτιμήσεις να είναι συνεπείς. Η βασική διαφορά μεταξύ των δύο προσεγγίσεων έγκειται στον τρόπο ορισμού των συνόλων κινδύνου και στο υπολογιστικό κόστος της διαδικασίας. Στην περίπτωση μεγάλου δείγματος και σπάνιου γεγονότος, η NCC προσέγγιση είναι, υπολογιστικά, σημαντικά πιο αποδοτική.

2.3.2.3 Λογιστική Παλινδρόμηση σε ζευγοποιημένες παρατηρήσεις ασθενών - μαρτύρων

Η προσέγγιση του μοντέλου παλινδρόμησης NCC συνδέεται στενά με τη λογιστική παλινδρόμηση matched case – control. Στο πλαίσιο αυτό εισάγονται νέοι συμβολισμοί για τις τιμές των συμμεταβλητών. Για κάθε παρατήρηση της ομάδας case $i = 1, \dots, M$:

- $X_{i:0}$: τιμή της συνεχούς μεταβλητής για το άτομο case
- $X_{i:1}, \dots, X_{i:m}$: τιμές της ίδιας μεταβλητής για τα m άτομα της ομάδας control που έχουν αντιστοιχιστεί στο άτομο i.

Ανάλογος συμβολισμός χρησιμοποιείται και για τις κατηγορικές μεταβλητές Z. Ορίζεται επίσης η δείκτρια συνάρτηση:

$$D_{i:k} = I(\text{το άτομο k απέτυχε τη χρονική στιγμή } T_i \mid \text{το άτομο } k \text{ επιβίωσε μέχρι τη χρονική στιγμή } T_i)$$

όπου για $k = 0$ το άτομο αντιστοιχεί στο ίδιο το case άτομο, ενώ για $k = 1, .., m$ αντιστοιχεί σε άτομα της ομάδας control. Το αντίστοιχο μοντέλο λογιστικής παλινδρόμησης διατυπώνεται ως:

$$\text{logit } P(D_{i:k} = 1 \mid X_{i:k}, Z_{i:k}) = a_i + X_{i:k}^T \boldsymbol{\beta}_x + Z_{i:k}^T \boldsymbol{\beta}_z \,, για\ k = 0,1, \dots, m \qquad (2.5)$$

όπου $\boldsymbol{\beta} = (\boldsymbol{\beta}_x^T, \boldsymbol{\beta}_z^T)^T$ είναι το διάνυσμα των συντελεστών και α_i μια παράμετρος που εκφράζει το βασικό κίνδυνο για το i – οστό σύνολο αντιστοίχισης. Στο Θεώρημα 1 της εργασίας τους, οι Yi et al. (2021) δείχνουν ότι, στην περίπτωση σπάνιας ασθένειας, οι εκτιμήσεις των συντελεστών του λογιστικού μοντέλου είναι προσεγγιστικά ίδιες με εκείνες του μοντέλου NCC (η απόδειξη του Θεωρήματος 1 δίνεται στο Μέρος Α των Υποστηρικτικών Πληροφοριών). Η συνάρτηση πιθανοφάνειας του λογιστικού μοντέλου παλινδρόμησης (2.5) δίνεται από:

$$L_{logit}(\boldsymbol{\beta}) = \prod_{i=1}^{M} M\left(\tilde{X}_i, \tilde{Z}_i; \boldsymbol{\beta}_x, \boldsymbol{\beta}_z\right) \tag{2.6}$$

$$\text{όπου} \quad M\left(\tilde{X}_i, \tilde{Z}_i; \boldsymbol{\beta}_x, \boldsymbol{\beta}_z\right) = P\left(D_{i:0} = 1, D_{i:1} = \cdots = D_{i:m} = 0 \mid \tilde{X}_i, \tilde{Z}_i, \sum_{k=0}^{m} D_{i:k} = 1\right)$$

$$= \left[1 + \sum_{k=1}^{m} \exp\{(X_{i:k} - X_{i:0})^T \boldsymbol{\beta}_x + (Z_{i:k} - Z_{i:0})^T \boldsymbol{\beta}_z\}\right]^{-1}$$

και για σπάνιες ασθένειες αποδεικνύεται ότι η πιθανοφάνεια (2.6) είναι πανομοιότυπη με τη μερική πιθανοφάνεια (2.4) λόγω του ότι:

$$M\left(\tilde{X}_i, \tilde{Z}_i; \boldsymbol{\beta}_x, \boldsymbol{\beta}_z\right) = \frac{\exp(\boldsymbol{X}_{i:0}^T \boldsymbol{\beta}_x + \boldsymbol{Z}_{i:0}^T \boldsymbol{\beta}_z)}{\sum_{k=0}^{m} \exp(\boldsymbol{X}_{i:k}^T \boldsymbol{\beta}_x + \boldsymbol{Z}_{i:k}^T \boldsymbol{\beta}_z)} \tag{2.7}$$

Η παραπάνω ισοδυναμία μεταξύ του NCC μοντέλου και της matched case – control λογιστικής παλινδρόμησης επιτρέπει την αξιοποίηση υπολογιστικά αποδοτικών τεχνικών εκτίμησης, γεγονός που καθιστά εφικτή την ανάπτυξη διαδικασιών επιλογής μεταβλητών σε δεδομένα πολύ υψηλής διάστασης, ιδιαίτερα σε μελέτες επιβίωσης που αφορούν σπάνιες ασθένειες. Στην επόμενη υποενότητα παρουσιάζεται η διαδικασία feature screening που προτείνουν οι Yi et al. (2021).

2.3.2.4 Προτεινόμενη διαδικασία φιλτραρίσματος

Σκοπός της συγκεκριμένης τεχνικής είναι η ανάπτυξη μιας υπολογιστικά αποδοτικής μεθόδου, ικανής να εφαρμοστεί σε μελέτες σπάνιων ασθενειών όπου τόσο το μέγεθος του δείγματος n όσο και ο αριθμός των επεξηγηματικών μεταβλητών p είναι ιδιαίτερα μεγάλα. Η βασική ιδέα είναι να αξιοποιηθεί το προηγούμενο πλαίσιο για να εξεταστεί κάθε στοιχείο $X_{i:0}$ και $Z_{i:0}$ για τα άτομα της ομάδας case. Ορίζεται η εξής ποσότητα:

$$\widetilde{\boldsymbol{W}}_i = \left[\left\{\sum_{k=1}^{m} (\boldsymbol{X}_{i:k} - \boldsymbol{X}_{i:0})\right\}^T, \left\{\sum_{k=1}^{m} (\boldsymbol{Z}_{i:k} - \boldsymbol{Z}_{i:0})\right\}^T\right]^T$$

όπου $i = 1, \ldots, M$ και για $l = 1, \ldots, p$ τέτοιο ώστε $\widetilde{W}_{i(l)}$ το l −οστό στοιχείο του $\widetilde{\boldsymbol{W}}_i$. Επίσης, ορίζεται το οριακό (marginal) μοντέλο Cox με συνάρτηση κινδύνου

$$\lambda\left(t|\widetilde{W}_{i(l)}\right) = \lambda_{0l}(t) \exp\left\{\widetilde{W}_{i(l)} \beta_l^M\right\} \tag{2.8}$$

όπου $\lambda_{0l}(t)$ είναι η αναφορική συνάρτηση κινδύνου και β_l^M είναι ο συντελεστής παλινδρόμησης. Έστω, επίσης, το σύνολο $M = \{l \in \{1, \dots, p\}: \beta_l \neq 0\}$ όπου περιλαμβάνει τις πραγματικά ενεργές συμμεταβλητές του μοντέλου (2.2) και $p_0 = |M|$ το μέγεθος του συνόλου M. Στη συνέχεια, συνεπής με Fan και Song (2010), απαιτούν τα οριακά σήματα να φράσσονται κάτω από θετική σταθερά έτσι ώστε να είναι ισχυρότερα από τους στοχαστικούς θορύβους και να μπορεί να γίνει διάκριση των ενεργών συμμεταβλητών από τις ανενεργές. Αυτή η απαίτηση δίνεται από τον τύπο:

$$\min_{l \in M} |\beta_l^M| \geq O(M^{1/2-\kappa})$$

όπου $0 < \kappa < 1/2$ και υποδηλώνει ότι το ελάχιστο απόλυτο μέγεθος των συντελεστών είναι τουλάχιστον της τάξης $M^{\frac{1}{2}-\kappa}$. Η διαδικασία για την επιλογή των ενεργών μεταβλητών αποτελείται από δύο σημαντικά βήματα:

1) Η εκτίμηση του δείκτη $r_l = \hat{\beta}_l^M / \mathrm{se}(\hat{\beta}_l^M)$ όπου ο συντελεστής β_l^M και το τυπικό του σφάλμα $\mathrm{se}(\beta_l^M)$ εκτιμώνται από τη μεγιστοποίηση της κάτωθι συνάρτησης πιθανοφάνειας:

$$L_l(\beta_l^M) = \prod_{i=1}^{M} M\left(\widetilde{W}_{i(l)}; \beta_l^M\right) \quad \text{όπου}$$

$$M\left(\widetilde{W}_{i(l)}; \beta_l^M\right) = P\left(D_{i:0} = 1, D_{i:1} = \cdots = D_{i:m} = 0 \mid \widetilde{W}_{i(l)}, \sum_{k=0}^{m} D_{i:k} = 1\right)$$
$$= \left[1 + \exp\left\{\widetilde{W}_{i(l)} \beta_l^M\right\}\right]^{-1}$$

2) Η σύγκριση του εκτιμώμενου δείκτη r_l με ένα προκαθορισμένο κατώφλι γ. Οι Yi et al. (2021) προτείνουν τη μη συμπερίληψη της συμμεταβλητής l στο σύνολο των ενεργών συμμεταβλητών M όταν ισχύει:

$$|r_l| < \gamma$$

Αυτή η model – based διαδικασία διαλογής μειώνει το αρχικό πλήρες μοντέλο (2.2) των παραμέτρων παλινδρόμησης p σε ένα μοντέλο με $p_0 \ll p$ παραμέτρους και λειτουργεί ικανοποιητικά σε περιπτώσεις όπου τόσο το p όσο και το n είναι μεγάλα. Το τίμημα που καταβάλλεται για το χειρισμό της υπολογιστικής έντασης που προκαλείται από το μεγάλο μέγεθος του n είναι η απαίτηση της υπόθεσης των σπάνιων ασθενειών. Πέραν από την εφαρμογή της μεθόδου στο πλαίσιο του ημιπαραμετρικού μοντέλου κινδύνου του Cox, εξετάζουν και την περίπτωση κατά την οποία το μοντέλο του Cox δεν έχει προσδιορισθεί σωστά. Στην περίπτωση αυτή εφαρμόζεται ένα μοντέλο AFT. Μέσω προσομοιώσεων διαπιστώνεται ότι η λανθασμένη προδιαγραφή μοντέλου μπορεί να οδηγήσει σε μη αξιόπιστα αποτελέσματα. Ως τρόπος βελτίωσης της απόδοσης της μεθόδου προτείνεται η αύξηση της παραμέτρου m, ιδιαίτερα όταν πρόκειται για δεδομένα υψηλής διάστασης. Επιπλέον, όταν εξετάζονται δεδομένα μεγάλης κλίμακας (δηλαδή όταν τόσο το n όσο και το p είναι αρκετά μεγάλα, με $n \gg p$), τα αποτελέσματα που

προκύπτουν από την εφαρμογή του μοντέλου Cox και ενός κατάλληλα επιλεγμένου μοντέλου AFT εμφανίζονται αρκετά παρόμοια.

Οι Yi et al. (2021) παρουσιάζουν εκτενώς τα θεωρητικά αποτελέσματα της προτεινόμενης μεθόδου. Συγκεκριμένα, στο θεώρημα 2 αποδεικνύεται ότι, υπό κατάλληλες συνθήκες κανονικότητας, ο εκτιμητής $\hat{\beta}_l^M$ βρίσκεται κοντά στην πραγματική τιμή του συντελεστή β_{l0}^M. Στο θεώρημα 3, διατυπώνεται ένα ομοιόμορφο φράγμα για την ποσότητα $\left(\hat{\beta}_l^M - \beta_{l0}^M\right)$, το οποίο εξασφαλίζει ότι η απόκλιση του εκτιμητή από την πραγματική τιμή παραμένει περιορισμένη. Η ιδιότητα αυτή επιτρέπει τον προσδιορισμό ενός μικρού αριθμού πιθανών ενεργών συμμεταβλητών p_0 σε σχέση με την αρχική διάσταση p, εφόσον το κατώφλι γ επιλεγεί κατάλληλα. Στο θεώρημα 4, προσδιορίζεται μια μορφή για το κατώφλι $\gamma = \gamma_M = bM^{1/2-k}$ με $0 < \kappa < 1/2$, η οποία εξασφαλίζει ένα κατώτερο όριο πιθανότητας (δίνεται στο Θεώρημα 4 της εργασίας) για την επιτυχία της διαδικασίας διαλογής, ώστε να συμπεριληφθούν όλες οι πραγματικά ενεργές μεταβλητές. Τέλος, στο θεώρημα 5 αποδεικνύεται ότι η προτεινόμενη μέθοδος ικανοποιεί την ιδιότητα sure screening, διατηρώντας ταυτόχρονα ελεγχόμενο το ποσοστό ψευδώς επιλεγμένων μεταβλητών. Οι αποδείξεις των θεωρημάτων 1 – 5 παρατίθενται στις Υποστηρικτικές Πληροφορίες της εργασίας Yi et al. (2021).

2.3.3 SIS με χρήση του δείκτη RMST

2.3.3.1 Ορισμός RMST Δείκτη

Οι Chen et al. (2024) πρότειναν μια model – free διαδικασία διαλογής επεξηγηματικών μεταβλητών, ικανή να ανιχνεύει γραμμικές και μη, μονοτονικές και μη, τοπικές και μη συσχετίσεις με τη μεταβλητή απόκρισης, η οποία στην παρούσα περίπτωση είναι ο χρόνος επιβίωσης *T*. Η θεωρητική βάση της μεθόδου στηρίζεται στην ύπαρξη δεξιάς λογοκρισίας (βλ. ενότητα 1.3.1). Το κύριο μέτρο της διαδικασίας είναι ο περιορισμένος μέσος χρόνος επιβίωσης (restricted mean survival time – RMST), ο οποίος ορίζεται ως εξής:

$$RMST(\tau) = \int_0^\tau S(t)dt$$

όπου

- $\tau > 0$: ένα καθορισμένο χρονικό σημείο με $P(Y > \tau) > 0$,
- $Y = \min(T, C)$: παρατηρούμενος χρόνος,
- T : χρόνος επιβίωσης,
- C: χρόνος λογοκρισίας,
- $\Delta = I(T \leq C)$: δείκτης αποτυχίας (1 εάν η παρατήρηση δεν είναι λογοκριμένη),
- $S(t)$: s.f. στο χρόνο t

Το RMST είναι ιδιαίτερα χρήσιμο για την εκτίμηση διαφορών στο χρόνο επιβίωσης μεταξύ ομάδων όταν η υπόθεση αναλογικού κινδύνου (PH) δεν ισχύει, όπως σε διασταυρώσεις καμπυλών επιβίωσης ή βραχυπρόθεσμες επιδράσεις θεραπείας. Για τον εντοπισμό σχέσης μεταξύ μιας επεξηγηματικής μεταβλητής X_j και του χρόνου επιβίωσης T, ορίζονται δύο διαστρωματοποιημένα RMST:

$$RMST_j^1(\tau_j^1; x) = \int_0^{\tau_j^1} S(t|X_j \geq x)dt \quad \& \quad RMST_j^2(\tau_j^2; x) = \int_0^{\tau_j^2} S(t|X_j < x)dt$$

όπου x ανήκει στο πεδίο ορισμού της X_j και τ_j^1, τ_j^2 είναι χρόνοι περιορισμού τέτοιοι ώστε $P(Y > \tau_j^1|X_j \geq x) > 0$ και $P(Y > \tau_j^2|X_j < x) > 0$ αντίστοιχα. Η συνολική απόκλιση RMST για την j −οστή συμμεταβλητή ορίζεται ως:

$$d_j = d_{j1} + d_{j2} = \int_{X_j} \left|RMST_j^1(\tau_j^1; x) - RMST(\tau_j^1)\right| dF_j(x) + \int_{X_j} \left|RMST_j^2(\tau_j^2; x) - RMST(\tau_j^2)\right| dF_j(x)$$

με $F_j(x)$ είναι η c.d.f. της μεταβλητής X_j. Μεγάλες τιμές του d_j υποδεικνύουν επίδραση της X_j στο χρόνο επιβίωσης T, ενώ τιμές κοντά στο 0 υποδεικνύουν ανεξαρτησία. Η εκτίμηση αυτής της ποσότητας δίνεται από τον τύπο:

$$\hat{d}_j = \frac{1}{n}\sum_{i=1}^{n}\left[\left|\widehat{RMST}_j^1(\hat{\tau}_{ji}^1; X_{ji}) - \widehat{RMST}(\hat{\tau}_{ji}^1)\right| + \left|\widehat{RMST}_j^2(\hat{\tau}_{ji}^2; X_{ji}) - \widehat{RMST}(\hat{\tau}_{ji}^2)\right|\right]$$

όπου οι εκτιμήσεις RMST προκύπτουν από του Kaplan – Meier εκτιμητές

- $\widehat{RMST}_j^1(\tau; X_{ji}) = \int_0^\tau \hat{S}(t|X_j \geq X_{ji})dt$
- $\widehat{RMST}_j^2(\tau; X_{ji}) = \int_0^\tau \hat{S}(t|X_j < X_{ji})dt$
- $\widehat{RMST}(\tau) = \int_0^\tau \hat{S}(t)dt$

και υπό κατάλληλους περιορισμούς

- $\hat{\tau}_{ji}^1 = \max_{\{k: X_{jk} \geq X_{ji}, \Delta_\kappa = 1\}}\{Y_k\}$
- $\hat{\tau}_{ji}^2 = \max_{\{k: X_{jk} < X_{ji}, \Delta_\kappa = 1\}}\{Y_k\}$

Το $\hat{d}_j$ είναι αμετάβλητο σε οποιονδήποτε μονοτονικό μετασχηματισμό της μεταβλητής X_j. Το θεώρημα 1 της εργασίας παρέχει την ασυμπτωτική κατανομή του $\hat{d}_{j1}$ υπό τη μηδενική υπόθεση ότι $d_{j1} = 0$. Επιπλέον, οι Chen et al. (2024) προτείνουν μια εκτίμηση του δείκτη υπό απουσία λογοκρισίας με τη μορφή:

$$\hat{d}_j = \hat{d}_{j1} + \hat{d}_{j2} = \frac{1}{n}\sum_{k=1}^{n}\left|\frac{1}{\{i: X_{ji} \geq X_{jk}\}}\sum_{i=1}^{n} Y_i I(X_{ji} \geq X_{jk}) - \frac{1}{n}\sum_{i=1}^{n} Y_i\right| + \frac{1}{n}\sum_{k=1}^{n}\left|\frac{1}{\{i: X_{ji} < X_{jk}\}}\sum_{i=1}^{n} Y_i I(X_{ji} < X_{jk}) - \frac{1}{n}\sum_{i=1}^{n} Y_i\right| \quad (2.9)$$

2.3.3.2 Μέθοδος φιλτραρίσματος με χρήση RMST δείκτη

Ο δείκτης $\hat{d}_j$ χρησιμοποιείται για την αναγνώριση των μεταβλητών που επηρεάζουν το χρόνο επιβίωσης T. Έστω το σύνολο ενεργών μεταβλητών:

$$M_* = \{j: E(T|X) \text{ επηρεάζεται λειτουργικά από κάποια } X_j, j = 1, \dots, p_n\}$$

με μέγεθος μη αραιότητας $s_n = |M_*|$. Ο όρος «λειτουργικά» αναφέρεται στον έλεγχο επίδρασης όλων των τύπων εξάρτησης της X_j (γραμμική και μη, τοπική και μη, μονοτονική και μη) με το χρόνο επιβίωσης *T*. Η εκτίμηση του M_* δίνεται από:

$$\widehat{M} = \{j: \hat{d}_j \geq \gamma_n, j = 1, \dots, p_n\}$$

Στην πράξη, επιλέγεται ένα κατώφλι γ_n τέτοιο ώστε να συμπεριλαμβάνονται οι μεταβλητές επίδρασης, συνήθως $[n/\log(n)]$ στο πλήθος, ή για να ελεγχθεί ποσοστό ψευδώς ανακαλυφθεισών μεταβλητών $\alpha\%$. Η διαδικασία πληροί την ιδιότητα sure screening εάν $\gamma_n = 1/2\ c_0 n^{-\varphi}$, με $c_0 > 0$ και $0 < \varphi < 1/2$ υπό τις συνθήκες 1 – 3 των Chen et al. (2024):

1) Αξιόπιστη εκτίμηση της επίδρασης της X_j στο χρόνο επιβίωσης *T*.
2) Διαφορά μεταξύ σήματος και θορύβου (ranking consistency property).
3) Διασφάλιση ότι η διάταξη των στατιστικών δεικτών $\hat{d}_j$ δεν είναι πολύ μικρή, εξασφαλίζοντας τη sure screening property.

2.3.3.3 Επαναληπτική διαδικασία φιλτραρίσματος με χρήση RMST δείκτη

Η επαναληπτική διαδικασία εφαρμόζεται ως εξής:

1) Εκτέλεση της διαδικασίας της προηγούμενης υποενότητας για εντοπισμό των καλύτερων $[n/\log(n)]$ συμμεταβλητών ως υποψήφιο σύνολο A^1.
2) Επαναληπτικά, όσο $|A^\kappa| < q$ και $A^\kappa \neq A^{\kappa-1}$:
   i) Επιλογή μεταβλητών από το A^κ μέσω Lasso penalty (βλ. ενότητα 2.4) σε ένα γενικευμένο προσθετικό ημιπαραμετρικό μοντέλο Cox (GAM Cox) για απόκτηση μη μηδενικών συντελεστών (σύνολο A).
   ii) Εκτέλεση GAM Cox με τις μεταβλητές του A για υπολογισμό των deviance κατάλοιπων, που χρησιμοποιούνται ως μεταβλητή απόκρισης για νέα διαδικασία διαλογής μεταξύ των $\{1, \dots, p_n\} \backslash A$ με το δείκτη (2.9). Επιλέγονται οι καλύτερες μεταβλητές πλήθους $\min\{[n/\log(n)],\ q - |A|\}$, οπότε προκύπτει το σύνολο M.
   iii) Ορισμός $\kappa = \kappa + 1$ και $A^\kappa = A \cup M$.
3) Τελικό output: A^κ

Στις προσομοιώσεις της εργασίας, χρησιμοποιήθηκε $q = [n/\log(n)]$. Περισσότερες πληροφορίες σχετικά με την εφαρμογή του GAM Cox παρέχονται στην εργασία των Chen et al. (2024). Δύο βασικοί περιορισμοί που προκύπτουν από την επαναληπτική διαδικασία είναι η απουσία τυποποιημένου τρόπου προσδιορισμού του αριθμού των επιλεγμένων μεταβλητών και η πιθανή παράβλεψη συμμεταβλητών οι οποίες δεν εμφανίζουν σημαντική οριακή συσχέτιση με το χρόνο επιβίωσης, αλλά μπορεί να σχετίζονται σημαντικά σε συνδυασμό με άλλες μεταβλητές. Η προτεινόμενη μεθοδολογία αποτελεί μια ολοκληρωμένη προσέγγιση feature screening, ικανή να ανιχνεύει μεταβλητές με μη απαραίτητα γραμμική σχέση με το χρόνο επιβίωσης, παρέχοντας ταυτόχρονα ερμηνεύσιμο μέτρο για ασθενείς και

επαγγελματίες σε κλινικές δοκιμές, ενώ διασφαλίζει ισχυρά θεωρητικά χαρακτηριστικά όπως η ιδιότητα sure screening. Η επαναληπτική διαδικασία διαχειρίζεται αποτελεσματικά φαινόμενα υψηλής πολυσυγγραμμικότητας και ύπαρξης ακραίων τιμών στις επεξηγηματικές μεταβλητές. Επιπλέον, οι Chen et al. (2024) δείχνουν ότι η μέθοδος μπορεί να εφαρμοστεί πέρα από τυχαία δεξιά λογοκρισία και σε δεδομένα με τυχαία λογοκρισία σε διάστημα, όπως αποδεικνύεται μέσα από παραδείγματα εφαρμογής, ακόμη και όταν δεν συνοδεύεται από θεωρητική τεκμηρίωση.

Όλα τα παραπάνω αποτελούν το μεγαλύτερο μέρος των μεθόδων feature screening που έχουν προταθεί στη βιβλιογραφία μέχρι σήμερα. Παρόλα αυτά, εξακολουθεί να υπάρχει χώρος για περαιτέρω μελέτη και βελτίωση κάθε μεθόδου, είτε σε επίπεδο θεωρητικών αποτελεσμάτων είτε ως προς τον ακριβέστερο προσδιορισμό ποσοτήτων που παραμένουν ανοιχτές ή έχουν διατυπωθεί μόνο μερικώς. Ορισμένες από τις μεθόδους feature screening που παρουσιάσθηκαν προηγουμένως οδηγούν άμεσα σε ένα μικρό σύνολο μεταβλητών, οι οποίες φαίνεται να επηρεάζουν το χρόνο επιβίωσης. Ωστόσο, υπάρχουν και μέθοδοι που στοχεύουν κυρίως στη μείωση της διάστασης του αρχικού συνόλου μεταβλητών σε έναν επιθυμητό αριθμό, ώστε να καταστεί εφικτή η εφαρμογή ποινικοποιημένης παλινδρόμησης στο ημιπαραμετρικό μοντέλο κινδύνου του Cox. Τέτοιες μέθοδοι θα παρουσιασθούν στο υπόλοιπο μέρος του κεφαλαίου.

2.4 Ποινικοποιημένη Παλινδρόμηση στο ημιπαραμετρικό μοντέλο του Cox

Η ανάλυση δεδομένων επιβίωσης υψηλής διάστασης έχει οδηγήσει στην ανάπτυξη μεθόδων που αποσκοπούν ταυτόχρονα στην επιλογή σημαντικών μεταβλητών, στη σταθεροποίηση των εκτιμητών και στη βελτίωση της ικανότητας πρόβλεψης. Στο πλαίσιο αυτό, η διαδικασία εντοπισμού των πραγματικά σημαντικών επεξηγηματικών μεταβλητών που επηρεάζουν το χρόνο επιβίωσης βασίζεται συνήθως σε δύο διαδοχικά στάδια, το feature screening και το regularization – penalization, όπως έχει ήδη αναφερθεί. Η βασική ιδέα του δεύτερου σταδίου είναι ότι η εκτίμηση στο μοντέλο Cox μπορεί να ελεγχθεί μέσω της προσθήκης ενός όρου ποινής στη λογαριθμική μερική συνάρτηση πιθανοφάνειας $l(\boldsymbol{\beta})$ (βλ. ενότητα 1.3). Με τον τρόπο αυτό, η διαδικασία εκτίμησης δεν περιορίζεται πλέον στην προσαρμογή των δεδομένων, αλλά ενσωματώνει και έναν μηχανισμό ελέγχου της πολυπλοκότητας του μοντέλου. Στην ουσία, διαφορετικές επιλογές ποινών αντιστοιχούν σε διαφορετικές στρατηγικές ισορροπίας μεταξύ ακρίβειας προσαρμογής και απλότητας του μοντέλου. Η πρώτη συστηματική προσέγγιση σε αυτό το πλαίσιο είναι η Ridge παλινδρόμηση, η οποία εισάγει ποινή L_2 από τους Hoerl & Kennard (1970), ενώ εφαρμογή στο μοντέλο του Cox έγινε από τους Verweij & Van Houwelingen (1994). Η μορφή της είναι:

$$\boldsymbol{Penalty} = \lambda\left|\left|\boldsymbol{\beta}\right|\right|_2^2 = \lambda\sum_{j=1}^{p}\beta_j^2\,, \qquad \lambda > 0$$

Η Ridge δεν οδηγεί σε επιλογή μεταβλητών, αλλά λειτουργεί σταθεροποιητικά, μειώνοντας τη διακύμανση των εκτιμητών και αντιμετωπίζοντας προβλήματα πολυσυγγραμμικότητας. Μ' αυτόν τον τρόπο, η Ridge μπορεί να θεωρηθεί ως μια μέθοδος που θυσιάζει την αραιότητα (sparsity) προς όφελος της σταθερότητας της εκτίμησης. Η ανάγκη για ταυτόχρονη εκτίμηση και επιλογή μεταβλητών οδήγησε στη Lasso, από τον Tibshirani (1996), η οποία εισάγει ποινή L_1:

$$\boldsymbol{Penalty} = \lambda\left|\left|\boldsymbol{\beta}\right|\right|_1 = \lambda\sum_{j=1}^{p}|\beta_j|, \qquad \lambda > 0$$

Η Lasso διαφοροποιείται θεμελιωδώς από τη Ridge, καθώς επιτρέπει τον ακριβή μηδενισμό συντελεστών και συνεπώς την επιλογή μεταβλητών. Ωστόσο, η ιδιότητα αυτή συνοδεύεται από πιθανή μεροληψία στις εκτιμήσεις, ειδικά για μεγάλους συντελεστές. Έτσι, η Lasso αποτελεί ένα πρώτο ουσιαστικό βήμα προς την έννοια της αραιότητας, εισάγοντας ταυτόχρονα το συμβιβασμό μεταξύ επιλογής και μεροληψίας. Η προσπάθεια βελτίωσης αυτών των περιορισμών οδήγησε σε πιο προσαρμοστικές εκδοχές, όπως η Adaptive Lasso των Zhang & Lu (2007), όπου η ποινή γράφεται ως:

$$\boldsymbol{Penalty} = \lambda\sum_{j=1}^{p}\tau_j|\beta_j|, \qquad \lambda > 0, \qquad \tau_j = 1/\left|\widetilde{\beta}_j\right|$$

Η εισαγωγή των βαρών επιτρέπει διαφοροποιημένη αντιμετώπιση των μεταβλητών, ενισχύοντας τις σημαντικές και αποδυναμώνοντας τις λιγότερο σχετικές. Με αυτόν τον τρόπο, η διαδικασία επιλογής γίνεται πιο προσαρμοστική στη δομή των δεδομένων, μειώνοντας τη συστηματική μεροληψία της κλασικής Lasso.

Η Elastic Net εισάγει έναν συνδυασμό των δύο βασικών ποινών L_1 και L_2, που προτάθηκε αρχικά από τους Zou & Hastie (2005) στη Γραμμική Παλινδρόμηση και ύστερα οι Park & Hastie (2007a) εφάρμοσαν το συγκεκριμένο penalty στα Γενικευμένα Γραμμικά Μοντέλα και πρότειναν σε δεύτερη φάση έναν αλγόριθμο εφαρμογής και στο αναλογικό μοντέλο κινδύνου του Cox (2007b). Η μορφή που δίνεται από τους Tibshirani et al. (2011) είναι:

$$\boldsymbol{Penalty} = \lambda P_a(\boldsymbol{\beta}) = \lambda\left(\alpha\sum_{i=1}^{p}|\beta_i| + \frac{1}{2}(1-a)\sum_{i=1}^{p}\beta_i^2\right), \qquad \lambda > 0, 0 < \alpha < 1$$

ενώ η εκτίμηση των συντελεστών $\boldsymbol{\beta}$ δίνεται από τον τύπο:

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}}\left(\frac{2}{n}\log(L(\boldsymbol{\beta})) - \boldsymbol{Penalty}\right)$$

Ο συνδυασμός αυτός επιτρέπει την ταυτόχρονη αξιοποίηση της αραιότητας της Lasso και της σταθερότητας της Ridge. Ιδιαίτερα σε περιπτώσεις έντονα συσχετισμένων μεταβλητών, η μέθοδος παρουσιάζει το λεγόμενο grouping effect. Με αυτόν τον τρόπο, το Elastic Net λειτουργεί ως γέφυρα μεταξύ της σταθεροποίησης και επιλογής μεταβλητών. Μετά το Elastic Net, είναι φυσικό να αναπτυχθούν επεκτάσεις που επιχειρούν να διατηρήσουν το πνεύμα του συνδυασμού L_1 και L_2, αλλά και να ενσωματώσουν επιπλέον πληροφορία για τη δομή ή τη συσχέτιση των μεταβλητών. Σε αυτή την κατεύθυνση, έχουν προταθεί πιο εξειδικευμένες παραλλαγές που επιχειρούν να επεκτείνουν το πλαίσιο της κανονικοποίησης σε πιο σύνθετες δομές δεδομένων. Μια τέτοια επέκταση είναι η Kernel Elastic Net των

Vinzamuri & Reddy (2013), η οποία εισάγει πληροφορία ομοιότητας μεταξύ μεταβλητών μέσω ενός kernel πίνακα και ορίζεται ως:

$$\boldsymbol{Penalty} = \lambda\left(\alpha\left|\left|\boldsymbol{\beta}\right|\right|_1 + (1-\alpha)\boldsymbol{\beta}^T K\boldsymbol{\beta}\right), \qquad \lambda \geq 0, \alpha \in (0,1)$$

όπου ο πίνακας K είναι ένας kernel (π.χ. RBF) που αποτυπώνει τη μη γραμμική συσχέτιση μεταξύ των επεξηγηματικών μεταβλητών. Η βασική ιδέα της προσέγγισης αυτής είναι ότι η ποινικοποίηση δεν βασίζεται πλέον μόνο στο μέγεθος των συντελεστών, αλλά και στη δομή ομοιότητας των ιδίων των μεταβλητών. Με αυτόν τον τρόπο, το μοντέλο είναι σε θέση να ενσωματώνει μη γραμμικές σχέσεις μεταξύ των χαρακτηριστικών, διατηρώντας παράλληλα την ιδιότητα της αραιότητας μέσω του L_1 όρου. Οι Vinzamuri & Reddy (2013) πρότειναν στην ίδια εργασία μια διαφορετική, αλλά συγγενής, προσέγγιση, τη λεγόμενη OSCAR (Octagonal Shrinkage and Clustering Algorithm for Regression). Η συγκεκριμένη ποινή εισάγει ταυτόχρονα επιλογή μεταβλητών και ομαδοποίηση συσχετισμένων παραμέτρων μέσω της ποινής:

$$\boldsymbol{Penalty} = \lambda_1\left|\left|\boldsymbol{\beta}\right|\right|_1 + \lambda_2\left|\left|T\boldsymbol{\beta}\right|\right|_1, \qquad \lambda_1 \geq 0, \lambda_2 \geq 0$$

Η βασική ιδέα είναι ότι η ποινή δεν επιβάλλει μόνο αραιότητα, αλλά και τείνει να εξισώνει συντελεστές που αντιστοιχούν σε συσχετισμένες μεταβλητές, δημιουργώντας έτσι ομάδες. Σε αντίθεση με το Elastic Net, το OSCAR δεν περιορίζεται στη σταθεροποίηση συσχετισμένων μεταβλητών, αλλά επιχειρεί ενεργά να τις «συγκεντρώσει» σε ομάδες με παρόμοιες τιμές συντελεστών. Αυτό οδηγεί σε μια μορφή ταυτόχρονης επιλογής και clustering των επεξηγηματικών μεταβλητών. Σε πιο πρόσφατες προσεγγίσεις, έχει δοθεί έμφαση όχι μόνο στη μορφή της ποινής, αλλά και στον τρόπο με τον οποίο αυτή μπορεί να επαναδιατυπωθεί μέσα στη διαδικασία εκτίμησης του μοντέλου Cox. Σε αυτή την κατεύθυνση, οι Mittal et al. (2014) πρότειναν μια τροποποιημένη εκδοχή των κλασικών Ridge και Lasso ποινών, με στόχο την καλύτερη προσαρμογή σε ultra – high dimensional δεδομένα επιβίωσης. Η βασική ιδέα τους είναι ότι η λογαριθμική μερική συνάρτηση πιθανοφάνειας μπορεί να γραφεί ως:

$$l_G(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left[\boldsymbol{\beta}^T \boldsymbol{x}_i - \log\left(\sum_{t \in R(y_i)} \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)\right)\right] - \boldsymbol{Penalty}$$

όπου η ποινή δεν εισάγεται απλώς ως L_1 ή L_2 όρος, αλλά με βάση την παρακάτω εναλλακτική διατύπωση. Για την περίπτωση του Lasso, η ποινή μπορεί να εκφραστεί ως:

$$\boldsymbol{Penalty}_{\boldsymbol{L_1}} = \sum_{j=1}^{p} \left(\log 2 - \log\left(\sqrt{\gamma_j}\right) + \sqrt{\gamma_j}\left|\beta_j\right|\right)$$

ενώ για τη Ridge:

$$\boldsymbol{Penalty}_{\boldsymbol{L_2}} = \sum_{j=1}^{p}\left(\log\sqrt{\tau_j} + \frac{1}{2}\log(2\pi) + \frac{\beta_j^2}{2\tau_j}\right)$$

Η ενδιαφέρουσα πτυχή αυτής της προσέγγισης είναι ότι, όταν οι παράμετροι γ_j και τ_j είναι κοινές για όλες τις μεταβλητές, ανακτώνται οι κλασικές μορφές Lasso και Ridge αντίστοιχα. Σε παρόμοια κατεύθυνση, η SCAD (Smoothly Clipped Absolute Deviation) ποινή προτάθηκε ως μη κυρτή εναλλακτική που στοχεύει στη διατήρηση της αραιότητας χωρίς την έντονη μεροληψία της Lasso:

$$\boldsymbol{Penalty} = \sum_{j=1}^{p} p_\lambda(|\beta_j|)$$

για το οποίο ισχύει ότι $p'_\lambda(|\beta_j|) = \lambda\, I(|\beta_j| \le \lambda) + \frac{(\alpha\lambda - |\beta_j|)_+}{(\alpha-1)} I(|\beta_j| > \lambda),\ \lambda > 0,\ \alpha > 2.$ Η βασική ιδέα της SCAD είναι ότι η συμπεριφορά της ποινής διαφοροποιείται ανάλογα με το μέγεθος των συντελεστών, προσεγγίζοντας τη Lasso για μικρές τιμές και αποφεύγοντας την υπερβολική συρρίκνωση για μεγάλες. Έτσι, επιτυγχάνεται ταυτόχρονα αραιότητα και σχεδόν αμερόληπτη εκτίμηση για σημαντικούς συντελεστές. Σε συνέχεια των μη κυρτών ποινών τύπου SCAD, έχουν προταθεί επεκτάσεις που επιχειρούν να ενσωματώσουν επιπλέον ευελιξία στο πλαίσιο του μοντέλου Cox, επιτρέποντας ταυτόχρονα πιο σύνθετες δομές στη σχέση μεταξύ μεταβλητών και χρόνου επιβίωσης. Σε αυτή την κατεύθυνση, οι Liam et al. (2014) πρότειναν μια εναλλακτική διατύπωση της SCAD ποινής στο πλαίσιο ενός επεκταμένου μοντέλου επιβίωσης, όπου η μερική ποινικοποιημένη λογαριθμική πιθανοφάνεια του μοντέλου γράφεται ως:

$$l(\boldsymbol{a}, \boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left\{ Z_i^T \boldsymbol{a} + X_i^T \boldsymbol{\beta} - log\left(\sum_{k=1}^{n} Y_k(T_i) exp[Z_i^T \boldsymbol{a} + X_i^T \boldsymbol{\beta}]\right)\right\} - n\sum_{j=1}^{p} p_\lambda(|\beta_j|)$$

με παράγωγο ποινής της μορφής:

$$p'_\lambda(|\beta_j|) = \lambda\, I(|\beta_j| \le \lambda) + \frac{(c\lambda - |\beta_j|)_+}{(c-1)} I(|\beta_j| > \lambda), \qquad c > 2$$

Η βασική διαφοροποίηση της προσέγγισης αυτής έγκειται στο γεγονός ότι το μοντέλο επεκτείνεται ώστε να περιλαμβάνει και μη παραμετρικό μέρος, επιτρέποντας μεγαλύτερη ευελιξία στη μορφή της σχέσης μεταξύ συμμεταβλητών και χρόνου επιβίωσης. Με αυτόν τον τρόπο, η SCAD δεν χρησιμοποιείται πλέον αποκλειστικά ως εργαλείο επιλογής μεταβλητών, αλλά ενσωματώνει επιλογή μεταβλητών και εκτίμηση μη γραμμικών συναρτήσεων. Η εκτίμηση των μη παραμετρικών όρων πραγματοποιείται μέσω spline – based προσεγγίσεων, όπως B – splines, επιτρέποντας την προσέγγιση ομαλών συναρτήσεων με περιορισμένο αριθμό βάσεων. Έτσι, η μέθοδος αποκτά ταυτόχρονα ιδιότητες αραιότητας και ευελιξίας. Η συγκεκριμένη προσέγγιση μπορεί να θεωρηθεί ως επέκταση των κλασικών SCAD μοντέλων σε ένα πλαίσιο όπου η δομή του μοντέλου δεν είναι αποκλειστικά γραμμική, αλλά περιλαμβάνει και μη γραμμικά, ομαλά στοιχεία. Η ανάγκη για πιο γενικές διατυπώσεις επιλογής μεταβλητών οδήγησε στον Survival Dantzig Selector των

Antoniadis et al. (2010), ο οποίος δεν βασίζεται άμεσα σε ποινικοποίηση της πιθανοφάνειας αλλά σε περιορισμούς της μορφής:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{\boldsymbol{p}}}\left|\left|\boldsymbol{\beta}\right|\right|_{\mathbf{1}} \;\; \upsilon\pi\acute{o} \;\; \left|\left|U(\boldsymbol{\beta})\right|\right|_{\infty} = \max_{j}\left|U_j\right| \leq \gamma, \qquad \gamma > 0$$

όπου $U(\boldsymbol{\beta})$ είναι ένα διάνυσμα $p \times 1$ διαστάσεων με τις παραγώγους των λογαρίθμων της μερικής συνάρτησης πιθανοφάνειας για κάθε επεξηγηματική μεταβλητή X_j. Η προσέγγιση αυτή οδηγεί σε αραιές λύσεις μέσω περιορισμών στο score function του μοντέλου. Σε αντίθεση με τις κλασικές μεθόδους ποινικοποίησης, η εκτίμηση εδώ διατυπώνεται ως πρόβλημα περιορισμένης βελτιστοποίησης, προσφέροντας μια διαφορετική οπτική στο ίδιο στατιστικό πρόβλημα. Μια διαφορετική κατεύθυνση προσφέρουν οι Liu et al. (2013), μέσω της γενικευμένης ποινής L_q:

$$\boldsymbol{Penalty} = \lambda\sum_{j=1}^{p}\left|\beta_j\right|^{q}, \qquad \lambda > 0, 0 < q < 1$$

Οι ποινές αυτές ενισχύουν την αραιότητα σε σχέση με τη Lasso και προσεγγίζουν τη L_0 κανονικοποίηση καθώς $q \to 0$. Έτσι, αποτελούν πιο «επιθετικές» μορφές επιλογής μεταβλητών, με ισχυρότερη τάση προς τις αραιές λύσεις. Οι Liu et al. (2013) επικεντρώνονται στην εργασία τους στην ποινή $L_{1/2}$. Όταν οι μεταβλητές έχουν φυσική ομαδοποίηση, οι group – based προσεγγίσεις των Kim et al. (2012) επιτρέπουν επιλογή σε επίπεδο J ομάδων μέσω περιορισμών της μορφής:

$$\sum_{j=1}^{J}\left|\left|\boldsymbol{\beta}_j\right|\right| \leq \lambda, \qquad \lambda > 0$$

Σε αυτή την περίπτωση, η επιλογή δεν γίνεται σε ατομικό επίπεδο αλλά σε επίπεδο ομάδων μεταβλητών. Αυτό είναι ιδιαίτερα χρήσιμο όταν οι επεξηγηματικές μεταβλητές έχουν φυσική ή εννοιολογική δομή. Σε ακόμη πιο εξειδικευμένες περιπτώσεις, η fused Lasso των Chaturvedi et al. (2014):

$$\boldsymbol{Penalty} = \lambda_1\sum_{j=1}^{p}\left|\beta_j\right| + \lambda_2\sum_{j=1}^{p-1}\left|\beta_{j+1} - \beta_j\right|, \qquad \lambda_1, \lambda_2 > 0$$

επιβάλλει ταυτόχρονα αραιότητα και ομαλότητα. Η προσθήκη του όρου διαφορών οδηγεί σε τμηματικά σταθερές εκτιμήσεις, ιδιαίτερα χρήσιμες όταν οι μεταβλητές έχουν φυσική διάταξη. Αντίστοιχα, οι graph – based προσεγγίσεις των Sun et al. (2014) ενσωματώνουν δομική πληροφορία μέσω πινάκων Laplacian:

$$\boldsymbol{Penalty} = \lambda_1\left|\left|\boldsymbol{\beta}\right|\right|_{1} + \frac{\lambda_2}{2}\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{L}\boldsymbol{\beta}, \qquad \lambda_1 \geq 0, \lambda_2 \geq 0$$

Οι μέθοδοι αυτές αξιοποιούν πληροφορία συσχέτισης μεταξύ μεταβλητών μέσω γραφήματος ή κάποιου δικτύου. Με αυτόν τον τρόπο, επιβάλλεται ομαλότητα στις εκτιμήσεις σύμφωνα με τη δομή των σχέσεων των επεξηγηματικών μεταβλητών. Τέλος, σε ένα πιο γενικό θεωρητικό επίπεδο, οι Bradic et al. (2011) πρότειναν ένα πλαίσιο μη κυρτών ποινών της μορφής:

$$\boldsymbol{Penalty} = n\sum_{j=1}^{p}p_{\lambda_n}\left(\left|\beta_j\right|\right), \qquad \lambda_n \geq 0$$

ενώ η πρώτη βασική συνθήκη, που ορίζουν οι Bradic et al. (2011), απαιτεί τον ορισμό της ποσότητας $\rho(t;\lambda_n) = p_{\lambda_n}(t)/\lambda_n$. Αυτή η ποσότητα θα πρέπει να είναι αύξουσα και κοίλη στο διάστημα $[0,+\infty)$ και να έχει μια συνεχή πρώτη παράγωγο $\rho'(t;\lambda_n)$ με $\rho'(0+;\lambda_n) > 0$. Επιπρόσθετα, $\rho'(t;\lambda_n)$ είναι αύξουσα σε $\lambda_n \geq 0$ και $\rho'(0+;\lambda_n)$ είναι ανεξάρτητο του λ_n. Οι υπόλοιπες συνθήκες διατυπώνονται λεπτομερώς στην εργασία των Bradic et al. (2011). Το πλαίσιο αυτό βασίζεται σε γενικές συνθήκες πάνω στη συνάρτηση ποινής, οι οποίες εξασφαλίζουν επιθυμητές ιδιότητες όπως consistency, oracle property και ασυμπτωτική κανονικότητα. Σε αντίθεση με συγκεκριμένες ποινές όπως η SCAD, εδώ δεν καθορίζεται μία συγκεκριμένη μορφή penalty, αλλά ένα σύνολο συνθηκών που προσδιορίζουν τη θεωρητικά «ορθή» συμπεριφορά του εκτιμητή. Έτσι, το πλαίσιο αυτό λειτουργεί ως θεωρητική γενίκευση πολλών μη κυρτών ποινών της βιβλιογραφίας.

Συνολικά, οι μέθοδοι ποινικοποίησης στο μοντέλο Cox συνιστούν ένα ενιαίο θεωρητικό πλαίσιο αντιμετώπισης προβλημάτων υψηλής διάστασης, όπου διαφορετικές επιλογές ποινής αντιστοιχούν σε διαφορετικούς συμβιβασμούς μεταξύ αραιότητας, σταθερότητας και δομικής πληροφορίας. Η ουσιαστική διαφοροποίηση μεταξύ των προσεγγίσεων δεν έγκειται μόνο στη μαθηματική μορφή των ποινών, αλλά κυρίως στον τρόπο με τον οποίο αυτές διαχειρίζονται το θεμελιώδη συμβιβασμό μεταξύ απλής ερμηνείας και στατιστικής απόδοσης. Φυσικά, σημαντική διαφορά αποτελεί και ο υπολογιστικός τρόπος ή αλγόριθμος που προτείνεται από κάθε ερευνητική εργασία. Παρά το μεγάλο αριθμό προτεινόμενων προσεγγίσεων, καμία μέθοδος δεν είναι καθολικά βέλτιστη σε όλα τα πλαίσια εφαρμογής. Η απόδοσή τους εξαρτάται από τα χαρακτηριστικά των δεδομένων, όπως η διάσταση, η συσχέτιση των μεταβλητών και η υποκείμενη δομή. Στις επόμενες υποενότητες παρουσιάζονται πιο σύγχρονες επεκτάσεις – penalties που επιχειρούν να συνδυάσουν πολλαπλές επιθυμητές ιδιότητες, όπως αραιότητα, σταθερότητα και αξιοποίηση δομικής πληροφορίας.

2.4.1 Νέες ποινές και αλγόριθμοι βασισμένοι στην ποινή Lasso

Η ποινή Lasso έχει αποτελέσει αντικείμενο εκτεταμένης μελέτης, καθώς συνοδεύεται από σημαντικά θεωρητικά αποτελέσματα που ευνοούν την αραιότητα των εκτιμήσεων. Η ιδιότητα αυτή καθίσταται ιδιαίτερα κρίσιμη σε προβλήματα υψηλής διάστασης, όπου επιδιώκεται ταυτόχρονα η μείωση της πολυπλοκότητας και η απλούστερη ερμηνεία των μοντέλων πρόβλεψης. Ωστόσο, παρά τα πλεονεκτήματά της, η Lasso παρουσιάζει και ορισμένους περιορισμούς, οι οποίοι έχουν οδηγήσει στην ανάπτυξη νεότερων επεκτάσεων. Στη συνέχεια παρουσιάζονται ενδεικτικά σύγχρονες τεχνικές που επιχειρούν να μετριάσουν αδυναμίες των κλασικών Lasso – type ποινών, προτείνοντας πιο ευέλικτες προσεγγίσεις.

Ένα κρίσιμο ζήτημα στις μεθόδους ποινικοποίησης αφορά τον προσδιορισμό της υπερπαραμέτρου λ, η οποία ελέγχει το βαθμό της κανονικοποίησης και επηρεάζει άμεσα το συμβιβασμό μεταξύ προσαρμογής και αραιότητας. Ο πλέον διαδεδομένος, αν και όχι μοναδικός, τρόπος επιλογής της βασίζεται σε διαδικασίες διασταυρούμενης επικύρωσης (K – fold cross – validation (CV)). Συγκεκριμένα, το σύνολο δεδομένων διαχωρίζεται σε K υποσύνολα, εκ των οποίων τα K – 1 χρησιμοποιούνται για την

εκπαίδευση του μοντέλου και το υπολειπόμενο υποσύνολο για την αξιολόγησή του. Η διαδικασία επαναλαμβάνεται Κ φορές, έτσι ώστε κάθε υποσύνολο να χρησιμοποιηθεί μια φορά ως σύνολο επικύρωσης – αξιολόγησης. Μέσω αυτής της επαναληπτικής διαδικασίας και με βάση ένα προκαθορισμένο κριτήριο βελτιστοποίησης που εξαρτάται από το μοντέλο που εφαρμόζει ο εκάστοτε αναλυτής – όπως στη Γραμμική Παλινδρόμηση, κριτήριο αποτελεί η ελαχιστοποίηση του μέσου τετραγωνικού σφάλματος – εκτιμάται η υπερπαράμετρος που ρυθμίζει τη βαρύτητα της ποινής που χρησιμοποιείται. Στην πράξη, το Κ λαμβάνει συνήθως τιμές μεταξύ 3 και 20, με συχνότερες επιλογές τις τιμές 5 και 10, χωρίς αυτό να αποτελεί αυστηρό κανόνα, καθώς ακραίες επιλογές ενδέχεται να επηρεάσουν δυσμενώς την αξιοπιστία της διαδικασίας. Τέλος, σε περιπτώσεις όπου το μοντέλο περιλαμβάνει περισσότερες από μία υπερπαραμέτρους – όπως συμβαίνει, για παράδειγμα, στο Elastic Net – η διαδικασία επιλογής καθίσταται πιο σύνθετη. Σε αυτά τα πλαίσια, έχουν προταθεί εξειδικευμένοι αλγόριθμοι και στρατηγικές αναζήτησης, οι οποίες επιτρέπουν την ταυτόχρονη και αποδοτική εκτίμηση των βέλτιστων συνδυασμών παραμέτρων, διατηρώντας τον έλεγχο του υπολογιστικού κόστους και της στατιστικής απόδοσης.

2.4.1.1 Improved Adaptive Lasso

Οι He et al. (2019) πρότειναν μια εναλλακτική προσέγγιση της ποινής Lasso, την Improved Adaptive Lasso, βασισμένη σε μεγάλο βαθμό στην ιδέα των Zhang & Lu (2007) οι οποίοι εισήγαγαν την ποινή Adaptive Lasso ως μια σταθμισμένη εκδοχή της κλασικής Lasso. Η βασική ιδέα συνίσταται στην αξιοποίηση αρχικών εκτιμήσεων των συντελεστών, προκειμένου να κατασκευαστούν κατάλληλα βάρη που διαφοροποιούν το βαθμό συρρίκνωσης μεταξύ των μεταβλητών. Στο πρώτο στάδιο της μεθοδολογίας, εφαρμόζεται η κλασική Lasso στο πλαίσιο του μοντέλου Cox, οδηγώντας στις εκτιμήσεις:

$$\widehat{\boldsymbol{\beta}}_{Lasso}(\lambda^*) = \underset{\beta}{\operatorname{argmax}} \left\{ \frac{1}{n} l_n(\boldsymbol{\beta}) - \lambda^* \left|\left|\boldsymbol{\beta}\right|\right|_1 \right\}$$

όπου:

- $l_n(\boldsymbol{\beta})$: είναι ο λογάριθμος της μερικής πιθανοφάνειας του μοντέλου Cox που αναλύθηκε στην ενότητα 1.3
- λ^* είναι η tuning parameter που καθορίζεται μέσω K – fold CV. Σύμφωνα με τους Simon et al. (2011), η εκτίμηση για την καλή προσαρμογή για την k – οστή κατανομή και μια παράμετρο κανονικοποίησης λ είναι

$$\widehat{CV}_{Lasso,k}(\lambda) = l_n\left(\widehat{\boldsymbol{\beta}}_{Lasso,-k}(\lambda)\right) - l_{n,-k}\left(\widehat{\boldsymbol{\beta}}_{Lasso,-k}(\lambda)\right)$$

όπου:

- $l_{n,-k}$ είναι η λογαριθμική μερική πιθανοφάνεια εξαιρουμένου του μέρους k των δεδομένων και
- $\widehat{\boldsymbol{\beta}}_{lasso,-k}(\lambda)$ είναι το $\boldsymbol{\beta}$ που προκύπτει από τα δεδομένα εκπαίδευσης (τα άλλα μέρη K −1) χρησιμοποιώντας λ:

$$\widehat{\boldsymbol{\beta}}_{Lasso,-k}(\lambda) = \underset{\beta}{\operatorname{argmax}} \left\{ \frac{1}{n} l_{n,-k}(\boldsymbol{\beta}) - \lambda \left|\left|\boldsymbol{\beta}\right|\right|_1 \right\}$$

Επαναλαμβάνοντας αυτή τη διαδικασία $k = 1, \dots, K$ φορές προκύπτει ότι:

$$\lambda^* = \underset{\lambda}{\text{arg}min} \sum_{k=1}^{K} \widehat{CV}_{Lasso,k}(\lambda)$$

Σε δεύτερο βήμα, με βάση την επιλεγμένη τιμή λ^* ελέγχονται ποιες μεταβλητές εμφανίζουν μηδενικούς συντελεστές και απορρίπτονται από το επόμενο βήμα. Έστω ότι το πλήθος των μη μηδενικών συντελεστών είναι $q < p$. Έπειτα, εφαρμόζεται η adaptive lasso διαδικασία των Zhang & Lu (2007) με τη διαφορά ότι ως βάρη, στην περίπτωση αυτή, ορίζονται οι τιμές των q μεταβλητών από το σύνολο $\widehat{\boldsymbol{\beta}}_{Lasso}(\lambda^*)$. Δηλαδή, οι νέοι συντελεστές δίνονται από τον τύπο:

$$\widehat{\boldsymbol{\beta}}_{adapt}(\gamma^*) = \underset{\beta}{\text{argmax}} \left\{ \frac{1}{n} l_n(\boldsymbol{\beta}) - \gamma^* \sum_{j=1}^{q} \frac{|\beta_j|}{|\hat{\beta}_{Lasso,j}(\lambda^*)|} \right\} \tag{2.10}$$

Με παρόμοιο τρόπο με το λ^* υπολογίζεται και το γ^*:

$$\gamma^* = \underset{\gamma}{\text{arg}min} \sum_{k=1}^{K} \widehat{CV}_{naive,k}(\gamma) \quad \text{όπου}$$

$$\widehat{CV}_{naive,k}(\lambda) = l_n\left(\widehat{\boldsymbol{\beta}}_{naive,-k}(\gamma)\right) - l_{n,-k}\left(\widehat{\boldsymbol{\beta}}_{naive,-k}(\gamma)\right)$$

και σ' αυτό το στάδιο, οι εκτιμήσεις των $\widehat{\boldsymbol{\beta}}_{naive,-k}(\gamma)$ προκύπτουν από την επίλυση του παρακάτω τύπου:

$$\widehat{\boldsymbol{\beta}}_{naive,-k}(\gamma^*) = \underset{\beta}{\text{argmax}} \left\{ \frac{1}{n} l_{n,-k}(\boldsymbol{\beta}) - \gamma \sum_{j=1}^{q} \frac{|\beta_j|}{|\hat{\beta}_{Lasso,j}(\lambda^*)|} \right\}$$

Ο δεύτερος αλγόριθμος (improved cross – validation) που προτάθηκε από τους He et al. (2019) διαφέρει από τον προηγούμενο όσον αφορά την εκτίμηση της υπερπαραμέτρου γ^*. Συγκεκριμένα, πρότειναν:

$$\gamma^* = \underset{\gamma}{\text{arg}min} \sum_{k=1}^{K} \widehat{CV}_{adapt,k}(\gamma)$$

με $\widehat{CV}_{adapt,k}(\lambda) = l_n\left(\widehat{\boldsymbol{\beta}}_{adapt,-k}(\gamma)\right) - l_{n,-k}\left(\widehat{\boldsymbol{\beta}}_{adapt,-k}(\gamma)\right)$ και σ' αυτό το στάδιο, οι εκτιμήσεις των $\widehat{\boldsymbol{\beta}}_{adapt,-k}(\gamma)$ προκύπτουν από την επίλυση του παρακάτω τύπου:

$$\widehat{\boldsymbol{\beta}}_{adapt,-k}(\gamma^*) = \underset{\beta}{\text{argmax}} \left\{ \frac{1}{n} l_{n,-k}(\boldsymbol{\beta}) - \gamma \sum_{j=1}^{q} \frac{|\beta_j|}{|\hat{\beta}_{-k,j}(\lambda^*_{-k})|} \right\}$$

όπου $\hat{\beta}_{-k,j}(\lambda^*_{-k})$ εκτιμάται ως ακολούθως:

Χωρίζεται το dataset, για $k = 1, \dots, K$ σε σύνολο εκπαίδευσης (training set) – $(K-1)$ στο πλήθος – για ακόμη μία φορά και εφαρμόζεται K–fold sub-CV. Τότε

$$\widehat{\boldsymbol{\beta}}_{-k}(\lambda^*_{-k}) = \underset{\beta}{argmax} \left\{ \frac{1}{n} l_{n,-k}(\boldsymbol{\beta}) - \lambda^*_{-k} \left|\left|\boldsymbol{\beta}\right|\right|_1 \right\}$$

με λ^*_{-k} επιλέγεται ώστε να ελαχιστοποιεί το άθροισμα όλων των στατιστικών στοιχείων καλής προσαρμογής μέσω της υποδιασταυρούμενης επικύρωσης (sub – CV) στο k – οστό σύνολο εκπαίδευσης.

Τέλος, στη συγκεκριμένη εργασία προτάθηκε και μια προσαρμοσμένη Elastic Net ποινή τέτοια ώστε αρχικά να εκτιμά τα

$$\widetilde{\boldsymbol{\beta}}_{EN} = \underset{\boldsymbol{\beta}}{argmax}\left\{\frac{1}{n}l_n(\boldsymbol{\beta}) - \lambda\left(\alpha\sum_{j=1}^{p}|\beta_j| + \frac{1}{2}(1-a)\sum_{j=1}^{p}\beta_j^2\right)\right\}$$

τα οποία, στη συνέχεια, χρησιμοποιούνται ως βάρη για το lasso μέρος της συγκεκριμένης ποινής:

$$\widetilde{\boldsymbol{\beta}}_{adapt} = \underset{\boldsymbol{\beta}}{argmax}\left\{\frac{1}{n}l_n(\boldsymbol{\beta}) - \lambda\left(\alpha\sum_{j=1}^{p}\frac{|\beta_j|}{|\tilde{\beta}_{EN,j}|} + \frac{1}{2}(1-a)\sum_{j=1}^{p}\beta_j^2\right)\right\}$$

Οι He et al. (2019) στην εργασία τους πρότειναν μια υπολογιστικά βολική και απλή διαδικασία, κατάλληλη για high dimensional καταστάσεις καθώς και για περιπτώσεις στρωματοποίησης και επαναλαμβανόμενων συμβάντων. Επίσης, διατηρεί χαμηλά τα επίπεδα των ψευδώς ανακαλύψεων, περιορίζοντας παράλληλα τις ψευδώς αρνητικές αναλογίες, παρουσιάζοντας έτσι βελτιωμένη απόδοση σε συγκεκριμένα σενάρια, σύμφωνα με τα αποτελέσματα της μελέτης.

2.4.1.2 Debiased Lasso

Οι Xia et al. (2023) πρότειναν μια σύγχρονη επέκταση της ποινής Lasso στο πλαίσιο του μοντέλου Cox, την Debiased Lasso, με βασικό στόχο την αντιμετώπιση της μεροληψίας που χαρακτηρίζει τους κλασικούς Lasso εκτιμητές. Η συμβολή τους έγκειται στην κατασκευή αμερόληπτων εκτιμητών των συντελεστών παλινδρόμησης, καθώς και στη δυνατότητα εξαγωγής διαστημάτων εμπιστοσύνης με ονομαστικές πιθανότητες κάλυψης. Με τον τρόπο αυτό καθίσταται εφικτή η στατιστική συμπερασματολογία μέσω ελέγχων υποθέσεων, ακόμη και σε περιβάλλοντα υψηλής διάστασης. Για τη διατύπωση της μεθοδολογίας εισάγεται ένα σύνολο συμβολισμών. Έστω $\boldsymbol{x} = (x_1, \dots, x_r)^T \in \mathbb{R}^r$ *με* $\boldsymbol{x}^{\otimes 0} = 1, \boldsymbol{x}^{\otimes 1} = \boldsymbol{x}, \boldsymbol{x}^{\otimes 2} = \boldsymbol{x}\boldsymbol{x}^T$. Η ℓ_q – νόρμα ορίζεται ως $||\boldsymbol{x}||_q = \left(\sum_{j=1}^{r}|x_j|^q\right)^{1/q}, q \geq 1$ ενώ για $q = 0$ ισχύει $||\boldsymbol{x}||_0 = \sum_{j=1}^{r} I(x_j \neq 0)$. Για έναν πίνακα $A = (\alpha_{ij}) \in \mathbb{R}^{m\times r}$, η επαγόμενη νόρμα δίνεται από

$$||A||_{q_1,q_2} = \sup_{\boldsymbol{x}\in\mathbb{R}^r, \boldsymbol{x}\neq 0}\left(\frac{||A\boldsymbol{x}||_{q_2}}{||\boldsymbol{x}||_{q_1}}\right), q_1, q_2 \geq 1$$

με ειδικές περιπτώσεις τις $||A||_{1,1} = \max_{1\leq j\leq r}\sum_{i=1}^{m}|a_{ij}|$, $||A||_{2,2} = \sigma_{max}(A)$ (η μεγαλύτερη μοναδική τιμή του A) και $||A||_{\infty,\infty} = \max_{1\leq j\leq r}\sum_{j=1}^{r}|a_{ij}|$. Η μέγιστη νόρμα ανά στοιχείο συμβολίζεται ως $||A||_{\infty} = \max_{i,j}|a_{ij}|$. Επιπλέον, για δύο θετικές ακολουθίες $\{d_n\}$ *και* $\{g_n\}$, ορίζεται $d_n \asymp g_n$ όταν ο λόγος τους φράσσεται άνω και κάτω από θετικές σταθερές. Στο πλαίσιο αυτό, θεωρείται το μοντέλο Cox με συνάρτηση κινδύνου:

$$h(t|X) = h_0(t)\exp(X^{\boldsymbol{T}}\boldsymbol{\beta}^0)$$

όπου $X = (\boldsymbol{X^{(1)}}, \dots, \boldsymbol{X^{(p)}}) \in \mathbb{R}^p$ είναι το διάνυσμα συμμεταβλητών και $\boldsymbol{\beta}^0$ το πραγματικό διάνυσμα παραμέτρων. Όταν η διάσταση p είναι σταθερή, η εκτίμηση μπορεί να πραγματοποιηθεί μέσω της κλασικής μεγιστοποίησης της μερικής πιθανοφάνειας. Αντιθέτως, σε καθεστώτα όπου $p \to \infty$, προτιμάται η χρήση του Lasso εκτιμητή:

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{arg\min}\left\{l_n(\boldsymbol{\beta}) + \lambda_n\left|\left|\boldsymbol{\beta}\right|\right|_1\right\}, \lambda_n > 0$$

Οι Xia et al. (2023) διατυπώνουν επίσης ρητές εκφράσεις για τη συνάρτηση σκορ (πρώτη παράγωγος) και τον πίνακα πληροφορίας (δεύτερη παράγωγος) της λογαριθμικής μερικής συνάρτησης πιθανοφάνειας, βασισμένες σε ποσότητες της μορφής

$$\hat{\mu}_r(t;\boldsymbol{\beta}) = \frac{1}{n}\sum_{j=1}^{n} I\big(Y_j \geq t\big)X_j^{\otimes r}\exp\big(X_j^T\boldsymbol{\beta}\big), \qquad r = 0,1,2$$

καθώς και στο σταθμισμένο μέσο διάνυσμα συμμεταβλητών

$$\hat{\eta}_n(t;\beta) = \frac{\hat{\mu}_1(t;\beta)}{\hat{\mu}_0(t;\beta)} = \frac{\sum_{j=1}^{n} I\big(Y_j \geq t\big)\exp\big(X_j^T\boldsymbol{\beta}\big)X_j}{\sum_{j=1}^{n} I\big(Y_j \geq t\big)\exp\big(X_j^T\boldsymbol{\beta}\big)}$$

Συγκεκριμένα δίνονται από τους τύπους

$$\dot{l_n}(\boldsymbol{\beta}) = -\frac{1}{n}\sum_{i=1}^{n}\delta_i\left\{X_i - \frac{\hat{\mu}_1(Y_i;\boldsymbol{\beta})}{\hat{\mu}_0(Y_i;\boldsymbol{\beta})}\right\}, \qquad \ddot{l_n}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\delta_i\left\{\frac{\hat{\mu}_2(Y_i;\boldsymbol{\beta})}{\hat{\mu}_0(Y_i;\boldsymbol{\beta})} - \left[\frac{\hat{\mu}_1(Y_i;\boldsymbol{\beta})}{\hat{\mu}_0(Y_i;\boldsymbol{\beta})}\right]^{\otimes 2}\right\}$$

Το βασικό μειονέκτημα της ποινής Lasso είναι η μεροληψία των εκτιμητών, ιδίως για μεγάλες τιμές της παραμέτρου ποινής. Για την αντιμετώπιση του προβλήματος αυτού, προτείνεται ένας διορθωμένος εκτιμητής της μορφής

$$\widehat{\boldsymbol{b}} = \big(\hat{b}_1, \dots, \hat{b}_p\big)^T = \widehat{\boldsymbol{\beta}} - \hat{\theta}\dot{l_n}\big(\widehat{\boldsymbol{\beta}}\big)$$

όπου ο όρος $-\hat{\theta}\dot{l_n}\big(\widehat{\boldsymbol{\beta}}\big)$ λειτουργεί ως διόρθωση πόλωσης, και $\hat{\theta}$ αποτελεί εκτίμηση του αντίστροφου πίνακα πληροφορίας. Η εκτίμηση του Θ, $\hat{\Theta}$, πραγματοποιείται μέσω επίλυσης ενός προβλήματος βελτιστοποίησης για κάθε στήλη του, σύμφωνα με την προσέγγιση των Javanmard & Montanari (2014):

$$\min\left\{\boldsymbol{m}^T\hat{\Sigma}\boldsymbol{m} : \boldsymbol{m} \in \mathbb{R}^p, \left|\left|\hat{\Sigma}\boldsymbol{m} - e_j\right|\right|_\infty \leq \gamma_n\right\}, j = 1, \dots, p$$

όπου $\gamma_n \geq 0$ είναι παράμετρος συντονισμού, e_j το μοναδιαίο διάνυσμα και

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\delta_i\big\{\boldsymbol{X}_i - \widehat{\boldsymbol{\eta}}_n\big(Y_i;\widehat{\boldsymbol{\beta}}\big)\big\}^{\otimes 2} = \frac{1}{n}\sum_{i=1}^{n}\delta_i\big\{\boldsymbol{X}_i - \widehat{\boldsymbol{\eta}}_n\big(Y_i;\widehat{\boldsymbol{\beta}}\big)\big\}\big\{\boldsymbol{X}_i - \widehat{\boldsymbol{\eta}}_n\big(Y_i;\widehat{\boldsymbol{\beta}}\big)\big\}^T$$

είναι ένας εκτιμητής του πίνακα πληροφορίας. Η προτεινόμενη μεθοδολογία επιτρέπει την αποφυγή άμεσης αντιστροφής του πίνακα πληροφορίας, γεγονός που μειώνει την υπολογιστική επιβάρυνση και καθιστά τη διαδικασία εφαρμόσιμη σε προβλήματα υψηλής διάστασης χωρίς αυστηρές υποθέσεις αραιότητας. Ωστόσο, η επιλογή της υπερπαραμέτρου γ_n είναι κρίσιμη για τη σταθερότητα της λύσης. Τα

θεωρητικά αποτελέσματα ισχύουν κυρίως στο καθεστώς «αποκλίνον p, μεγάλο n», ενώ σε περιπτώσεις όπου $\mathrm{p} > \mathrm{n}$, η συμπεριφορά του εκτιμητή ενδέχεται να διαφοροποιείται, απαιτώντας προσεκτική ερμηνεία των αποτελεσμάτων.

2.4.2 Νέες ποινές και αλγόριθμοι βασισμένοι στην ποινή Ridge

Η ποινή Ridge δεν αποτέλεσε σημείο αναφοράς στον ίδιο βαθμό με τη Lasso, κυρίως λόγω της έλλειψης ιδιότητας αραιότητας και των συναφών θεωρητικών πλεονεκτημάτων που αυτή συνεπάγεται. Ωστόσο, διατηρεί θεμελιώδη ρόλο σε προβλήματα πολυσυγγραμμικότητας, ενώ τα τελευταία χρόνια έχουν προταθεί επεκτάσεις που επιχειρούν να συνδυάσουν τη σταθερότητα της Ridge με ιδιότητες επιλογής μεταβλητών.

2.4.2.1 Broken Adaptive Ridge

Οι Kawaguchi et al. (2020) πρότειναν μια επαναληπτική μεθοδολογία αραιής παλινδρόμησης στο πλαίσιο του μοντέλου Cox, βασισμένη στην ποινή Ridge, η οποία ενσωματώνει ιδιότητες αραιότητας μέσω μιας διαδικασίας επανασταθμισμένης εκτίμησης. Η προσέγγιση αυτή, γνωστή ως Broken Adaptive Ridge (BAR), συνδέεται εννοιολογικά με προσεγγίσεις τύπου ℓ_0, επιτυγχάνοντας παράλληλα ευνοϊκές θεωρητικές ιδιότητες, όπως η αραιότητα, η oracle property και η ομαδοποίηση υψηλά συσχετισμένων μεταβλητών. Η διαδικασία ξεκινά με την εκτίμηση ενός μοντέλου Cox με την κλασική ποινή Ridge:

$$\widehat{\boldsymbol{\beta}}^{(0)} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ -2l_n(\boldsymbol{\beta}) + \xi_n \sum_{j=1}^{p} \beta_j^2 \right\}$$

Στη συνέχεια, εφαρμόζεται μια επαναληπτική διαδικασία επανασταθμισμένης Ridge, όπου σε κάθε βήμα οι εκτιμήσεις ενημερώνονται ως:

$$\widehat{\boldsymbol{\beta}}^{(k)} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ -2l_n(\boldsymbol{\beta}) + \lambda_n \sum_{j=1}^{p} \frac{\beta_j^2}{\left(\widehat{\beta}_j^{(k-1)}\right)^2} \right\}, k \geq 1$$

όπου ξ_n *και* λ_n είναι μη αρνητικές παράμετροι ρύθμισης ποινής. Ο τελικός εκτιμητής BAR ορίζεται ως το όριο της παραπάνω ακολουθίας:

$$\widehat{\boldsymbol{\beta}} = \lim_{k \to \infty} \widehat{\boldsymbol{\beta}}^{(k)}$$

Η συγκεκριμένη επανασταθμισμένη δομή επιτρέπει την επιβολή ισχυρότερης ποινικοποίησης σε συντελεστές μικρού μεγέθους, οδηγώντας τους πρακτικά στο μηδέν, ενώ ταυτόχρονα διατηρεί σταθερές τις εκτιμήσεις των σημαντικών μεταβλητών. Με τον τρόπο αυτό επιτυγχάνεται μια μορφή προσαρμοστικής αραιότητας, παρά το γεγονός ότι η αρχική ποινή είναι τύπου Ridge. Για τον υπολογισμό των εκτιμητών σε κάθε επανάληψη προτείνεται η χρήση της μεθόδου Newton – Raphson, η οποία συμβάλλει στην αποφυγή αριθμητικών προβλημάτων, όπως η αριθμητική υπερχείλιση. Ωστόσο, σε περιβάλλοντα υπερυψηλής διάστασης (ultra high – dimensional data), η εφαρμογή της μεθόδου ενδέχεται να καταστεί υπολογιστικά απαιτητική, λόγω αυξημένου κόστους, απαιτήσεων μνήμης και πιθανής αριθμητικής αστάθειας. Από τα αποτελέσματα προσομοιώσεων προκύπτει ότι η

απόδοση της μεθόδου επηρεάζεται κυρίως από την παράμετρο λ_n, ενώ η παράμετρος ξ_n έχει περιορισμένη επίδραση στο τελικό αποτέλεσμα. Επιπλέον, επισημαίνεται ότι η επιλογή της παραμέτρου μέσω διασταυρούμενης επικύρωσης ενδέχεται να οδηγήσει σε ασυμπτωτική υπερπροσαρμογή με θετική πιθανότητα, ενώ η χρήση κριτηρίων πληροφορίας, όπως το BIC, εμφανίζει συνέπεια ως προς την ανάκτηση του πραγματικού μοντέλου. Τέλος, προτείνεται μια γενίκευση της μεθόδου μέσω τροποποίησης των βαρών της επαναληπτικής διαδικασίας

$$1/\left(\hat{\beta}_j^{(k-1)}\right)^{2-d} \ , d \in [0,1]$$

Η παράμετρος d λειτουργεί ως μηχανισμός ελέγχου της αραιότητας. Συγκεκριμένα, καθώς το d αυξάνεται προς τη μονάδα, οι εκτιμήσεις καθίστανται λιγότερο αραιές, με ταυτόχρονη αύξηση των ψευδώς θετικών και της μεροληψίας, ιδίως σε περιπτώσεις μεγάλης διάστασης. Αντίθετα, μειώνει τον αριθμό των ψευδώς αρνητικών, γεγονός που αναδεικνύει το ρόλο του d ως παραμέτρου εξισορρόπησης μεταξύ ευαισθησίας και ειδικότητας.

2.4.2.2 Βελτιωμένη συρρίκνωση τύπου Ridge

Η Seyedeh Z. Aghamohammadi (2025) προτείνει μια οικογένεια τεχνικών κανονικοποίησης, στις οποίες, με βασικό μηχανισμό ποινικοποίησης τύπου L_2, επιτυγχάνεται η απόρριψη στατιστικά ασθενών επεξηγηματικών μεταβλητών. Μέσω προσομοιώσεων τεκμηριώνεται ότι η προτεινόμενη μεθοδολογία εφαρμόζεται αποτελεσματικά τόσο σε προβλήματα χαμηλής όσο και υψηλής διάστασης, παρουσιάζοντας αυξημένη απόδοση ιδίως σε περιβάλλοντα έντονης συσχέτισης μεταξύ των επεξηγηματικών μεταβλητών. Η διαδικασία βελτιωμένης συρρίκνωσης τύπου Ridge (Improved Ridge-type shrinkage) διατυπώνεται ως επαναληπτικό σχήμα εκτίμησης των συντελεστών παλινδρόμησης στο πλαίσιο του μοντέλου Cox. Η αρχική εκτίμηση προκύπτει από την επίλυση της εξίσωσης $U(\boldsymbol{\beta}) = 0$, όπου η συνάρτηση ορίζεται ως:

$$U(\boldsymbol{\beta}) = \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} d_i \left(\boldsymbol{x}'_{(i)} - \frac{\sum_{j \in R(t_{(i)})} \boldsymbol{x}_j \exp(\boldsymbol{x}_j' \boldsymbol{\beta})}{\sum_{j \in R(t_{(i)})} \exp(\boldsymbol{x}_j' \boldsymbol{\beta})} \right)$$

με $R(t) = \{i : T_i \geq t\}$ το σύνολο κινδύνου και το $\boldsymbol{x}_j$ το διάνυσμα παρατηρήσεων του $j-$οστού ατόμου. Η επίλυση του ανωτέρου συστήματος πραγματοποιείται μέσω επαναληπτικής διαδικασίας Newton – Raphson οδηγώντας σε αρχική εκτίμηση $\widehat{\boldsymbol{\beta}}^{(\mathbf{0})}$. Η ακρίβεια του αρχικού σημείου εκκίνησης είναι κρίσιμη, καθώς σημαντικές αποκλίσεις δύνανται να επηρεάσουν αρνητικά τη σύγκλιση του αλγορίθμου. Σε κάθε επανάληψη m, η επικαιροποιημένη εκτίμηση δίνεται από:

$$\widehat{\boldsymbol{\beta}}^{(m+1)} = \widehat{\boldsymbol{\beta}}^{(m)} + \left(X' \widehat{W}^{(m)} X \right)^{-\mathbf{1}} U(\widehat{\boldsymbol{\beta}}^{(m)})$$

όπου $\widehat{W}^{(m)}$ είναι διαγώνιος πίνακας διαστάσεων $n \times n$ με στοιχεία $w_{ii} = \exp\left(\boldsymbol{x}_j' \widehat{\beta}^{\,(m)}\right) H_0(t_i)$ και $H_0(t) = \sum_{s \leq t} h_0(s)$. Η διαδικασία τερματίζεται όταν ικανοποιείται το κριτήριο σύγκλισης

$$\left|\left| \widehat{\boldsymbol{\beta}}^{(m+1)} - \widehat{\boldsymbol{\beta}}^{(m)} \right|\right| \leq \varepsilon$$

οπότε και προκύπτει ο εκτιμητής μεγίστης μερικής πιθανοφάνειας. Στο πλαίσιο κανονικοποίησης, προτείνεται ο μη περιορισμένος εκτιμητής κορυφογραμμής:

$$\widehat{\boldsymbol{\beta}}^{UR} = \left(X'\widehat{W}\,X + \lambda \boldsymbol{I_p}\right)^{-1} X'\widehat{W}\,X\,\widehat{\boldsymbol{\beta}}$$

όπου $\lambda > 0$ αποτελεί υπερπαράμετρο κανονικοποίησης, $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}^{(m+1)}$, $\widehat{W} = \widehat{W}^{(m+1)}$, $I_{\boldsymbol{p}}$: μοναδιαίος πίνακας $p-$ διαστάσεων. Ο εκτιμητής αυτός παρουσιάζει μειωμένο μέσο τετραγωνικό σφάλμα σε σχέση με τον εκτιμητή μεγίστης πιθανοφάνειας. Βασιζόμενη στο έργο των Hoerl & Kennard (1970), προτείνεται εκτίμηση της παραμέτρου λ μέσω της σχέσης $\hat{\lambda} = 1/\hat{\varphi}'\hat{\varphi}$, όπου $\hat{\varphi} = C'\widehat{\boldsymbol{\beta}}$ με C ο πίνακας ιδιοδιανυσμάτων του $X'\widehat{W}\,X$. Σε περιπτώσεις όπου διατίθεται εκ των προτέρων πληροφορία υπό τη μορφή γραμμικών περιορισμών $H_0: R\boldsymbol{\beta} = \boldsymbol{r}$ (με R είναι ένας γνωστός πίνακας διαστάσεων $p_2 \times p$ $(p_2 \leq p)$ και $\boldsymbol{r}$ ένα σταθερό (fixed) διάνυσμα – στήλη διαστάσεων $p_2 \times 1$ σταθερών τιμών στο $\mathbb{R}^{p_2}$) ο αντίστοιχος περιορισμένος εκτιμητής κορυφογραμμής δίνεται από:

$$\widehat{\boldsymbol{\beta}}^{RR} = \left(X'\widehat{W}\,X + \lambda I_p\right)^{-1} X'\widehat{W}\,X\,\widetilde{\boldsymbol{\beta}}$$

$$\text{όπου } \widetilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}} - \mathcal{I}^{-1}R'(R\mathcal{I}^{-1}\boldsymbol{R}')^{-1}\left(R\widehat{\boldsymbol{\beta}} - \boldsymbol{r}\right), \qquad \mathcal{I}^{-1} = \left(X'\widehat{W}\,X\right)^{-1}$$

Για τη σύγκριση των εκτιμητών εισάγεται έλεγχος τύπου Wald με στατιστική συνάρτηση:

$$\mathcal{L}_n = n\left(\boldsymbol{R}\widehat{\boldsymbol{\beta}}^{UR} - \boldsymbol{r}\right)'\left(\boldsymbol{R}\left(\frac{1}{n}\mathcal{I}\right)^{-1}\boldsymbol{R}'\right)^{-1}\left(\boldsymbol{R}\widehat{\boldsymbol{\beta}}^{UR} - \boldsymbol{r}\right)$$

η οποία ασυμπτωτικά ακολουθεί κατανομή $\chi^2_{p_2}$. Η τελική εκτίμηση προκύπτει ως γραμμικός συνδυασμός των δύο εκτιμητών:

$$\widehat{\boldsymbol{\beta}}^{SR} = \widehat{\boldsymbol{\beta}}^{RR} + \left(1 - \frac{c}{\mathcal{L}_n}\right)\left(\widehat{\boldsymbol{\beta}}^{UR} - \widehat{\boldsymbol{\beta}}^{RR}\right)$$

όπου $c = p_2 - 2$ με $p_2 \geq 3$. Ωστόσο, για μικρές τιμές του $\mathcal{L}_n$, παρατηρείται υπερβολική συρρίκνωση του εκτιμητή UR προς το RR. Για την αντιμετώπιση του φαινομένου αυτού, προτείνεται η περικομμένη εκδοχή:

$$\widehat{\boldsymbol{\beta}}^{PSR} = \widehat{\boldsymbol{\beta}}^{RR} + \left(1 - \frac{c}{\mathcal{L}_n}\right)^{+}\left(\widehat{\boldsymbol{\beta}}^{UR} - \widehat{\boldsymbol{\beta}}^{RR}\right)$$

όπου $\left(1 - \frac{c}{\mathcal{L}_n}\right)^{+} = \max\left\{0, 1 - \frac{c}{\mathcal{L}_n}\right\}$. Ο εκτιμητής αυτός συμβάλλει στον περιορισμό της υπερβολικής συρρίκνωσης, ενισχύοντας τη σταθερότητα και την ευκολότερη ερμηνεία των εκτιμήσεων.

2.4.3 Άλλες σύγχρονες τεχνικές ποινικοποίησης

2.4.3.1 Ποινικοποιημένο μοντέλο αναλογικών κινδύνων Cox με ομαδοποιημένους προβλεπτικούς παράγοντες

Οι Dang et al. (2021) αξιοποίησαν τρεις θεμελιώδεις ποινές της στατιστικής, τις Lasso, SCAD και MCP, και στηριζόμενοι σε αυτές ανέπτυξαν δύο εκτενείς μεθοδολογικές προσεγγίσεις: τις μη επικαλυπτόμενες και τις επικαλυπτόμενες ομάδες. Η βασική ιδέα της εργασίας τους έγκειται στη δομημένη διαχείριση επεξηγηματικών μεταβλητών μέσω ομαδοποίησης, με στόχο την ενίσχυση της αποδοτικότητας των εκτιμητών και την απλούστερη ερμηνεία. Κρίσιμο στοιχείο για την επιτυχή εφαρμογή των μεθόδων αυτών αποτελεί η εκ των προτέρων γνώση ή μια εύλογη υπόθεση σχετικά με την ύπαρξη υποομάδων στο σύνολο δεδομένων.

Στην περίπτωση των μη επικαλυπτόμενων ομάδων, οι ομάδες μεταβλητών συγκροτούν ξένα μεταξύ τους σύνολα, δηλαδή κάθε μεταβλητή ανήκει σε μια και μόνο ομάδα. Υποτίθεται η ύπαρξη J τέτοιων διακριτών ομάδων, οι οποίες καλύπτουν το σύνολο των επεξηγηματικών μεταβλητών. Στο πλαίσιο αυτό, οι ποινές ενσωματώνονται στη λογαριθμική μερική πιθανοφάνεια του μοντέλου Cox, επιβάλλοντας αραιότητα στους συντελεστές παλινδρόμησης. Συγκεκριμένα, ορίζονται:

Group Lasso

$$Penalty = P_\lambda(\beta) = \sum_{j=1}^{J} \lambda_j \left|\left|\beta_j\right|\right| \quad \text{όπου } \lambda_j = \lambda\sqrt{p_j} \quad \mu\varepsilon\ j = 1, \dots, J$$

και p_j: ο αριθμός των μεταβλητών στην ομάδα j.

Group SCAD

$$Penalty = P_{\lambda,\gamma}(\beta) = \sum_{j=1}^{J} S_{\lambda,\gamma}\left(\left|\left|\beta_j\right|\right|\right) \quad \text{όπου}$$

$$S_{\lambda,\gamma}\left(\left|\left|\beta_j\right|\right|\right) = \begin{cases} \lambda_j \left|\left|\beta_j\right|\right|, & \varepsilon\acute{\alpha}\nu\ \left|\left|\beta_j\right|\right| \le \lambda_j \\ \dfrac{\gamma\lambda_j \left|\left|\beta_j\right|\right| - 0.5\left(\left|\left|\beta_j\right|\right|^2 + \lambda_j^2\right)}{\gamma - 1}, & \varepsilon\acute{\alpha}\nu\ \lambda_j < \left|\left|\beta_j\right|\right| \le \gamma\lambda_j \\ \dfrac{\lambda_j^2(\gamma^2 - 1)}{2(\gamma - 1)}, & \varepsilon\acute{\alpha}\nu\ \left|\left|\beta_j\right|\right| > \gamma\lambda_j \end{cases}$$

Group MCP

$$Penalty = P_{\lambda,\gamma}(\beta) = \sum_{j=1}^{J} M_{\lambda,\gamma}\left(\left|\left|\beta_j\right|\right|\right) \quad \text{όπου}$$

$$M_{\lambda,\gamma}\left(\left|\left|\beta_j\right|\right|\right) = \begin{cases} \lambda_j \left|\left|\beta_j\right|\right| - \dfrac{\left|\left|\beta_j\right|\right|^2}{2\gamma}, & \varepsilon\acute{\alpha}\nu\ \left|\left|\beta_j\right|\right| \le \gamma\lambda_j \\ \dfrac{1}{2}\gamma\lambda_j^2, & \varepsilon\acute{\alpha}\nu\ \left|\left|\beta_j\right|\right| > \gamma\lambda_j \end{cases}$$

όπου $\left|\left|\cdot\right|\right|$ είναι η ευκλείδεια νόρμα ενός διανύσματος. Οι Dang et al. (2021) προσεγγίζοντας την επίλυση κάθε επιμέρους προβλήματος με τον αλγόριθμο ελαχιστοποίησης μείζονος πιθανοφάνειας (MM) των Lange et al. (2000) και Hunter & Lange (2004) διατύπωσαν κλειστούς τύπους για τις εκτιμήσεις των συντελεστών $\boldsymbol{\beta}$**:**

Group Lasso

$$Q(\beta|\beta^*) = l(\beta^*) + l'(\beta^*)^T(\beta - \beta^*) + \frac{\tau}{2}(\beta - \beta^*)^T(\beta - \beta^*) + \sum_j \lambda_j \left|\left|\beta_j\right|\right|$$

όπου $Q_j'(\beta)$ είναι η μερική παράγωγος της $Q(\beta)$ για την ομάδα j και ορίζεται ως εξής:

$$Q_j'(\beta) = l_j'(\beta^*) + \tau\left(\beta_j - \beta_j^*\right) + \begin{cases} \lambda_j \dfrac{\beta_j}{\left|\left|\beta_j\right|\right|}, & \text{εάν } \beta_j \neq 0 \\ \lambda_j \left|\left|\boldsymbol{v}\right|\right|, & \text{εάν } \beta_j = 0 \end{cases}$$

Η εκτίμηση των $\widehat{\beta_j}$ της παραπάνω σχέσης έχει την ακόλουθη έκφραση κλειστής μορφής:

$$\widehat{\beta_j} = \left(1 - \frac{\lambda_j}{\tau\left|\left|\boldsymbol{r}\right|\right|}\right)_+ \boldsymbol{r}$$

Group SCAD

$$Q(\beta) = l(\beta^*) + l'(\beta^*)^T(\beta - \beta^*) + \frac{\tau}{2}(\beta - \beta^*)^T(\beta - \beta^*) + \sum_j S_{\lambda,\beta}\left(\left|\left|\beta_j\right|\right|\right)$$

Η βέλτιστη λύση χαρακτηρίζεται από την εξίσωση μερικής παραγώγου.

Εάν $\left|\left|\beta_j\right|\right| \leq \lambda_j$ τότε

$$\begin{cases} l_j'(\beta^*) + \tau\left(\beta_j - \beta_j^*\right) + \lambda_j \dfrac{\beta_j}{\left|\left|\beta_j\right|\right|} = 0\,, & \text{εάν } \beta_j \neq 0 \\ l_j'(\beta^*) - \tau\beta_j^* + \lambda_j \left|\left|\boldsymbol{v}\right|\right| = 0, & \text{εάν } \beta_j = 0 \end{cases}$$

Εάν $\lambda_j < \left|\left|\beta_j\right|\right| \leq \gamma\lambda_j$ τότε

$$l_j'(\beta^*) + \tau\left(\beta_j - \beta_j^*\right) + \frac{\gamma\lambda_j \dfrac{\beta_j}{\left|\left|\beta_j\right|\right|} - \beta_j}{\gamma - 1} = 0$$

Εάν $\left|\left|\beta_j\right|\right| > \gamma\lambda_j$ τότε

$$l_j'(\beta^*) + \tau\left(\beta_j - \beta_j^*\right) = 0$$

Ο τελικός κλειστός τύπος των εκτιμήσεων σε αυτό το penalty είναι:

$$\widehat{\beta_j} = \begin{cases} \left(1 - \dfrac{\lambda_j}{\tau\left|\left|\boldsymbol{r}\right|\right|}\right)_+ \boldsymbol{r}, & \text{εάν } \left|\left|\boldsymbol{r}\right|\right| \leq \left(\lambda_j + \dfrac{\lambda_j}{\tau}\right) \\ \dfrac{\tau(\gamma-1)}{\tau(\gamma-1)-1}\left(1 - \dfrac{\gamma\lambda_j}{\tau(\gamma-1)\left|\left|\boldsymbol{r}\right|\right|}\right)\boldsymbol{r}, & \text{εάν } \begin{cases} \tau(\gamma-1) - 1 > 0 \\ \left(\lambda_j + \dfrac{\lambda_j}{\tau}\right) < \left|\left|\boldsymbol{r}\right|\right| \leq \gamma\lambda_j \end{cases} \\ \boldsymbol{r}, & \text{εάν } \left|\left|\boldsymbol{r}\right|\right| > \gamma\lambda_j \end{cases}$$

Group MCP

$$Q(\beta) = l(\beta^*) + l'(\beta^*)^T(\beta - \beta^*) + \frac{\tau}{2}(\beta - \beta^*)^T(\beta - \beta^*) + \sum_j M_{\lambda,\beta}\left(\left|\left|\beta_j\right|\right|\right)$$

Εάν $\left|\left|\beta_j\right|\right| \leq \gamma\lambda_j$ τότε

$$\begin{cases} l_j'(\beta^*) + \tau\left(\beta_j - \beta_j^*\right) + \lambda_j \dfrac{\beta_j}{\left|\left|\beta_j\right|\right|} - \dfrac{1}{\gamma}\beta_j = 0\,, & \varepsilon\acute{\alpha}\nu\ \beta_j \neq 0 \\ l_j'(\beta^*) - \tau\left(\beta_j^*\right) + \lambda_j\left|\left|\boldsymbol{v}\right|\right| = 0, & \varepsilon\acute{\alpha}\nu\ \beta_j = 0 \end{cases}$$

Εάν $\left|\left|\beta_j\right|\right| > \gamma\lambda_j$ τότε

$$l_j'(\beta^*) + \tau\left(\beta_j - \beta_j^*\right) = 0$$

Ο τελικός κλειστός τύπος των εκτιμήσεων σε αυτό το penalty είναι:

$$\widehat{\beta_j} = \begin{cases} \dfrac{\tau\gamma}{\tau\gamma - 1}\left(1 - \dfrac{\lambda_j}{\tau\left|\left|\boldsymbol{r}\right|\right|}\right)_+ \boldsymbol{r}, & \varepsilon\acute{\alpha}\nu \left|\left|\boldsymbol{r}\right|\right| \leq \gamma\lambda_j, \tau\gamma - 1 > 0 \\ \boldsymbol{r}, & \varepsilon\acute{\alpha}\nu\ \left|\left|\boldsymbol{r}\right|\right| > \gamma\lambda_j \end{cases}$$

Και στις τρεις περιπτώσεις ορίζεται ένα οποιοδήποτε διάνυσμα $\boldsymbol{v}$ που ικανοποιεί τη σχέση $\left|\left|\boldsymbol{v}\right|\right| \leq 1$ ενώ $\boldsymbol{r} = \beta_j^* - \frac{l_j'(\beta^*)}{\tau}$ και $(x)_+ = \max\{x, 0\}$.

Στη δεύτερη τεχνική που εξετάστηκε στη συγκεκριμένη εργασία, η μεθοδολογία είναι πιο περίπλοκη και γι' αυτό δεν δίνονται κλειστοί τύποι των εκτιμητών των συντελεστών της παλινδρόμησης Cox. Στην προκειμένη περίπτωση, θεωρείται ότι κάποιες μεταβλητές μπορεί να ανήκουν σε περισσότερες από μία ομάδες και γι' αυτό αναφέρονται σε επικαλυπτόμενες ομάδες, δηλαδή οι ομάδες δεν αποτελούν ξένα σύνολα. Έστω ότι υπάρχουν $G-$ ομάδες, με p_g να συμβολίζεται ο αριθμός των επεξηγηματικών μεταβλητών στην $g-$οστή ομάδα. Οι ποινές που ορίστηκαν έχουν την εξής μορφή:

Group Lasso

$$Penalty = \Omega_\lambda(\beta) = \sum_{g=1}^{G} \lambda_g \left|\left|\beta_g\right|\right| \ \ \acute{o}\pi o\upsilon\ \lambda_g = \lambda\sqrt{p_g} \ \ \mu\varepsilon\ g = 1, \dots, G$$

$$\kappa\alpha\iota\ p_g\text{: } o\ \alpha\rho\iota\theta\mu\acute{o}\varsigma\ \tau\omega\nu\ \mu\varepsilon\tau\alpha\beta\lambda\eta\tau\acute{\omega}\nu\ \sigma\tau\eta\nu\ o\mu\acute{\alpha}\delta\alpha\ g$$

Group SCAD

$$Penalty = \Omega_{\lambda,\gamma}(\beta) = \sum_{g=1}^{G} S_{\lambda,\gamma}\left(\left|\left|\beta_g\right|\right|\right) \ \ \acute{o}\pi o\upsilon$$

$$S_{\lambda,\gamma}\left(\left||\beta_g|\right|\right)=\begin{cases}\lambda_g\left||\beta_g|\right|, & \text{εάν } \left||\beta_g|\right|\le\lambda_g\\ \dfrac{\gamma\lambda_g\left||\beta_g|\right|-0.5\left(\left||\beta_g|\right|^2+\lambda_g^2\right)}{\gamma-1}, & \text{εάν } \lambda_g<\left||\beta_g|\right|\le\gamma\lambda_g\\ \dfrac{\lambda_g^2(\gamma^2-1)}{2(\gamma-1)}, & \text{εάν } \left||\beta_g|\right|>\gamma\lambda_g\end{cases}$$

Group MCP

$$Penalty=\Omega_{\lambda,\gamma}(\beta)=\sum_{g=1}^{G}M_{\lambda,\gamma}\left(\left||\beta_g|\right|\right)\ \text{όπου}$$

$$M_{\lambda,\gamma}\left(\left||\beta_g|\right|\right)=\begin{cases}\lambda_g\left||\beta_g|\right|-\dfrac{\left||\beta_g|\right|^2}{2\gamma}, & \text{εάν } \left||\beta_g|\right|\le\gamma\lambda_g\\ \dfrac{1}{2}\gamma\lambda_g^2, & \text{εάν } \left||\beta_g|\right|>\gamma\lambda_g\end{cases}$$

Η εκτίμηση των $\boldsymbol{\beta}$ προκύπτει από ελαχιστοποίηση της παρακάτω ποσότητας

$$\mathcal{L}(\beta)=-\frac{1}{N}\log\big(L(\beta)\big)+Penalty$$

Οι Dang et al. (2021) προχωρούν και σε μία αναδιατύπωση της συνάρτησης $\mathcal{L}(\beta)$ έτσι ώστε να λάβει παρόμοια δομή με την περίπτωση των μη επικαλυπτόμενων ομάδων και να μπορεί να επιλυθεί με τον ίδιο αλγόριθμο που προτάθηκε παραπάνω. Στην συγκεκριμένη εργασία, οι Dang et al. (2021) παρουσιάζουν ορισμένα σημαντικά πλεονεκτήματα που διευκολύνουν υπολογιστικά και συμπερασματολογικά τον εκάστοτε αναλυτή. Αρχικά, σε σχέση με τα αρχικά penalties, έχουν αναπτυχθεί πιο σταθεροί και γρήγοροι αλγόριθμοι χωρίς το αρχικό βήμα ορθοκανονικοποίησης. Για το χειρισμό του πίνακα σχεδιασμού $\boldsymbol{X}$ χρησιμοποίησαν αλγορίθμους καθόδου ανά ομάδα, ενώ τον αλγόριθμο MM για τη μεγιστοποίηση. Απέδειξαν ότι ο συγκεκριμένος αλγόριθμος είναι γρήγορος και αποτελεσματικός.

Η ποινή group lasso που πρότειναν αποδίδει αραιότητα τόσο σε ομαδικό όσο και σε ατομικό επίπεδο. Ως μειονέκτημα επισημαίνεται ότι η group lasso τείνει στην επιλογή πιο σύνθετων μοντέλων, ενώ οι μη κυρτές ποινές, συμπεριλαμβανομένης της ομάδας SCAD και της ομάδας MCP, παρουσιάζουν πολλά υποσχόμενα αποτελέσματα επιλογής ομαδοποιημένων μεταβλητών με oracle properties. Τέλος, εντός αλγορίθμων χρειάστηκε σε όλα τα penalties ο ορισμός μιας παραμέτρου γ την οποία όρισαν $\gamma = 3.7$ για τη SCAD ομάδας, όπως προτείνεται από τους Fan & Li (2001) και $\gamma = 3$ για τη MCP ομάδας, όπως προτείνεται από τον Zhang (2010). Ωστόσο, προσαρμόζοντας τις τιμές γ, η group MCP προσεγγίζει την Group Lasso με $\gamma \to \infty$ και αντιστοίχως την group SCAD, ενώ ταυτόχρονα τονίζουν τη σημασία της κατάλληλης επιλογής της παραμέτρου γ.

2.4.3.2 Ποινικοποίηση των προβλεπτικών παραγόντων με χρήση γραφημάτων

Οι Xie et al. (2025) προτείνουν μια διαφορετική προσέγγιση σε σχέση με την υπάρχουσα βιβλιογραφία. Πιο συγκεκριμένα, η μεθοδολογία τους βασίζεται σε ένα μη κατευθυνόμενο και μη σταθμισμένο γράφημα $G = (V, E)$, το οποίο χρησιμοποιείται για την αποτύπωση της συσχέτισης μεταξύ των επεξηγηματικών μεταβλητών στο μοντέλο του Cox. Η συσχέτιση αυτή μπορεί να προκύπτει είτε από εκ των προτέρων διαθέσιμη πληροφορία του αναλυτή είτε από ένα προεκτιμημένο γράφημα πρόβλεψης που βασίζεται σε δεδομένα δείγματος. Πιο αναλυτικά, τα σύνολα V και E αντιστοιχούν στους κόμβους και στις άκρες του γραφήματος, αντίστοιχα. Για κάθε μεταβλητή i ορίζεται ένας κόμβος $i \in V$, ενώ υπάρχει μια άκρη $(i, j) \in E$ όταν οι μεταβλητές $i \, και \, j$ συσχετίζονται. Έστω E_G ο πίνακας που αναπαριστά το σύνολο των άκρων, όπου $E_G(i, j) = 1$ εάν $\{(i, j) \in E \, ή \, i = j\}$, και 0 διαφορετικά. Επιπλέον, ορίζεται το σύνολο γειτόνων του κόμβου i ως $N_i = \{j : E_G(i, j) = 1\}$ ενώ το πλήθος των στοιχείων του συμβολίζεται ως $d_i = |N_i|$. Στόχος της εργασίας είναι η εκτίμηση των συντελεστών παλινδρόμησης του μοντέλου Cox μέσω της εισαγωγής μιας ποινής που βασίζεται στο γράφημα G, το οποίο ενσωματώνει πρόσθετη πληροφορία για τον πίνακα σχεδιασμού X. Θεωρώντας ένα διάνυσμα μη αρνητικών βαρών $\tau := \{\tau_1, \dots, \tau_p\}$, η ποινή ορίζεται ως:

$$Penalty = \lambda \left|\left|\boldsymbol{\beta}\right|\right|_{G,\tau} = \lambda \times \min_{\substack{\sum_{k=1}^{p} V^{(k)} = \boldsymbol{\beta} \\ supp\left(V^{(k)}\right) \subseteq N_k}} \left\{ \sum_{k=1}^{p} \tau_k \left|\left|V^{(k)}\right|\right|_2 \right\}$$

όπου $V^{(k)}$ εκφράζει την επίδραση της $k-$οστής επεξηγηματικής μεταβλητής στο χρόνο επιβίωσης. Ειδικότερα, αν $V^{(k)} \neq 0$ υποδηλώνεται η ύπαρξη σύνδεσης μεταξύ της συγκεκριμένης μεταβλητής και των γειτονικών της κόμβων στο γράφημα G. Τα βάρη $\tau_k \geq 0$ λειτουργούν ως παράμετροι κανονικοποίησης που σταθμίζουν τη σημαντικότητα του όρου $\left|\left|V^{(k)}\right|\right|_2$. Η τελική μορφή της μερικής συνάρτησης πιθανοφάνειας, η οποία ελαχιστοποιείται για την εκτίμηση των παραμέτρων $\boldsymbol{\beta}$, δίνεται από:

$$\min_{\boldsymbol{\beta}, V^{(1)}, \dots, V^{(p)}} \left\{ -\frac{1}{n} \boldsymbol{l}(\boldsymbol{\beta}) + \lambda \sum_{k=1}^{p} \tau_k \left|\left|V^{(k)}\right|\right|_2 \right\}$$

υπό τους περιορισμούς $\sum_{k=1}^{p} V^{(k)} = \boldsymbol{\beta} \, και \, supp\left(V^{(k)}\right) \subseteq N_k$. Τέλος, στην εργασία τους οι Xie et al. (2025) προτείνουν μια αναδιατύπωση της παραπάνω συνάρτησης σε ένα μη περιορισμένο κυρτό πρόβλημα, με στόχο τη διευκόλυνση της διαδικασίας επίλυσης του προβλήματος ελαχιστοποίησης. Περισσότερες λεπτομέρειες παρουσιάζονται αναλυτικά στην αντίστοιχη εργασία.

3ο Κεφάλαιο
Τεχνικές Μηχανικής Μάθησης ως Μοντέλα Πρόβλεψης και Επιλογής Μεταβλητών

3.1 Εισαγωγή

Στο 2ο Κεφάλαιο παρουσιάστηκαν μέθοδοι επιλογής μεταβλητών που αποσκοπούν στον εντοπισμό ενός περιορισμένου συνόλου επεξηγηματικών μεταβλητών οι οποίες σχετίζονται ουσιαστικά με το χρόνο επιβίωσης (T). Σε πρώτο στάδιο, η μείωση της διάστασης επιτυγχάνεται μέσω διαδικασιών feature screening, είτε στο πλαίσιο του ημιπαραμετρικού μοντέλου Cox είτε με τη χρήση μεθόδων που δεν προϋποθέτουν συγκεκριμένη μορφή μοντελοποίησης της σχέσης μεταξύ των μεταβλητών και του χρόνου επιβίωσης. Στη συνέχεια, μέσω της εισαγωγής κατάλληλων όρων ποινής στη συνάρτηση μερικής πιθανοφάνειας, επιτυγχάνεται ταυτόχρονα η εκτίμηση των παραμέτρων και η επιλογή των σημαντικών επεξηγηματικών μεταβλητών, οδηγώντας σε πιο σταθερά και ερμηνεύσιμα μοντέλα σε συνθήκες υψηλής διάστασης. Ωστόσο, οι μέθοδοι που παρουσιάστηκαν στο 2ο Κεφάλαιο βασίζονται, κατά κύριο λόγο, στην υπόθεση μιας γραμμικής σχέσης μεταξύ των επεξηγηματικών μεταβλητών και της συνάρτησης κινδύνου, μέσω της ημιπαραμετρικής διατύπωσης του μοντέλου Cox. Συγκεκριμένα, η επίδραση των συμμεταβλητών εισέρχεται γραμμικά στη log – hazard συνάρτηση, γεγονός που συνεπάγεται μια προκαθορισμένη και σχετικά περιοριστική μορφή για τον τρόπο με τον οποίο οι μεταβλητές επηρεάζουν το χρόνο επιβίωσης.

Η διερεύνηση πιο σύνθετων σχέσεων, όπως μη γραμμικότητας ή αλληλεπιδράσεις μεταξύ μεταβλητών, είναι δυνατή κυρίως μέσω κατάλληλων μετασχηματισμών των αρχικών μεταβλητών από τον αναλυτή. Τέτοιες προσεγγίσεις δεν είναι πάντοτε εφικτές, καθώς προϋποθέτουν εκ των προτέρων γνώση της δομής των δεδομένων ή βασίζονται σε εμπειρικές επιλογές, οι οποίες ενδέχεται να μην αποτυπώνουν πλήρως την πραγματική πολυπλοκότητα του φαινομένου. Για τον λόγο αυτό, η αποτελεσματικότητα των παραδοσιακών προσεγγίσεων μπορεί να περιορίζεται σε περιπτώσεις όπου η σχέση μεταξύ συμμεταβλητών και χρόνου επιβίωσης αποκλίνει σημαντικά από τη γραμμική δομή που επιβάλλεται στη συνάρτηση κινδύνου. Επιπλέον, η ανάγκη για μεγαλύτερη ευελιξία στη μοντελοποίηση, ιδιαίτερα σε δεδομένα υψηλής διάστασης με πιθανές μη γραμμικές αλληλεπιδράσεις, οδηγεί στην υιοθέτηση πιο ευέλικτων μεθόδων πρόβλεψης. Στο πλαίσιο αυτό, στο παρόν κεφάλαιο εξετάζονται βασικές κατηγορίες μεθόδων μηχανικής μάθησης (machine learning – ML) για δεδομένα επιβίωσης. Ειδικότερα, παρουσιάζονται δέντρα επιβίωσης (survival trees), τυχαία δάση επιβίωσης (random survival forests), μέθοδοι ενίσχυσης (boosting) προσαρμοσμένες στο πλαίσιο της ανάλυσης επιβίωσης, καθώς και προσεγγίσεις βασισμένες σε διανύσματα υποστήριξης (support vector machines) για time–to–event δεδομένα επιβίωσης. Οι μέθοδοι αυτές αποτελούν

αντιπροσωπευτικά παραδείγματα ευέλικτων, μη παραμετρικών ή ημιπαραμετρικών μεθόδων πρόβλεψης, οι οποίες έχουν αποδειχθεί ιδιαίτερα αποτελεσματικές σε σύγχρονες εφαρμογές ιατρικής στατιστικής και βιοϊατρικής ανάλυσης.

3.2 Δέντρα Επιβίωσης

Τα Δέντρα Επιβίωσης οφείλουν την ονομασία τους στην ιεραρχική δομή τους, η οποία εκκινεί από έναν αρχικό κόμβο στην κορυφή και διακλαδίζεται προς τα κάτω. Ανάλογα με τις τιμές των μεταβλητών μιας νέας παρατήρησης, ακολουθείται μια συγκεκριμένη διαδρομή (path) που καταλήγει σε έναν τερματικό κόμβο. Στην περίπτωση αυτή, η τελική εκτίμηση αφορά την επιβίωση ενός ατόμου με βάση τα χαρακτηριστικά εκείνα που αξιολογήθηκαν ως σημαντικά από τον αλγόριθμο. Το συγκεκριμένο μοντέλο αποτελεί ένα ισχυρό εργαλείο πρόβλεψης, το οποίο διευκολύνει τη μελέτη και την έγκαιρη διαχείριση της υπό εξέταση έκβασης (π.χ. της κατάστασης ενός ασθενούς). Ο όρος «επιβίωση» προέρχεται απευθείας από το ευρύτερο γνωστικό αντικείμενο της Ανάλυσης Επιβίωσης.

Τα δέντρα επιβίωσης αποτελούν μια μη παραμετρική εναλλακτική προσέγγιση έναντι του ημιπαραμετρικού μοντέλου αναλογικών κινδύνων του Cox, καθώς και των AFT μοντέλων. Το βασικό τους πλεονέκτημα είναι ότι καταφέρνουν να εντοπίζουν πολύπλοκες αλληλεπιδράσεις μεταξύ των μεταβλητών με σχετικά απλό τρόπο, χωρίς να απαιτείται εκ των προτέρων γνώση ή καθοδήγηση της διαδικασίας. Επιπλέον, η ανάπτυξη ενός δέντρου επιτρέπει τη φυσική ομαδοποίηση των παρατηρήσεων με βάση τις επεξηγηματικές μεταβλητές. Η ιδέα της ανάπτυξης των δέντρων επιβίωσης ξεκίνησε να διατυπώνεται αρχικά από τους Ciampi et al. (1981) και Marubini et al. (1983), ωστόσο η πρώτη εργασία που ενσωμάτωσε όλα τα τυπικά χαρακτηριστικά ενός κλασικού δέντρου επιβίωσης παρουσιάσθηκε από τους Gordon and Olshen (1985). Κατά τα επόμενα έτη, πολλοί ερευνητές εισήγαγαν εναλλακτικά κριτήρια διαχωρισμού (splitting criteria) των κόμβων (nodes). Στον Πίνακα 3.1 παρουσιάζονται σε χρονική αλληλουχία τα κυριότερα κριτήρια διαχωρισμού. Σημειώνεται ότι οι βιβλιογραφικές αναφορές έως το έτος 2011 βασίζονται στην ανασκόπηση των Bou – Hamad et al. (2011), ενώ ο πίνακας έχει επεκταθεί έως το 2020 προκειμένου να συμπεριλάβει μεθοδολογικές εξελίξεις που αφορούν δεδομένα υψηλής διάστασης (high – dimensional data).

Συγγραφείς (Χρονολογία)	**Κριτήριο**
Gordon & Olshen (1985)	Μετρική Wasserstein μεταξύ της **s.f.** Kaplan – Meier και μιας **s.f.** που έχει μάζα το πολύ σε ένα πεπερασμένο σημείο
Ciampi et al. (1986)	Ελεγχοσυνάρτηση logrank για τη σύγκριση των δύο ομάδων που σχηματίζονται στους θυγατρικούς κόμβους
Ciampi et al. (1987)	Έλεγχος Λόγου Πιθανοφανειών (Likelihood Ratio Statistic - LRS) υπό ένα υποθετικό μοντέλο για τη μέτρηση ανομοιότητας μεταξύ των δύο θυγατρικών κόμβων

Συγγραφείς (Χρονολογία)	Κριτήριο
Ciampi et al. (1988) και Ciampi et al. (1989)	Ελεγχοσυνάρτηση logrank και Wilcoxon – Gehan ως μέτρα ανομοιότητας και της στατιστικής Kolmogorov – Smirnov για τη σύγκριση των καμπυλών επιβίωσης των δύο κόμβων
Segal (1988)	Ελεγχοσυνάρτηση Tarone – Ware δύο δειγμάτων για λογοκριμένα δεδομένα με κατάλληλη επιλογή βαρών
Therneau et al. (1990)	Κατάλοιπα martingale από ένα μηδενικό μοντέλο Cox
Leblanc & Crowley (1992)	Μέτρο απόκλισης κόμβου μεταξύ μιας κορεσμένης λογαριθμικής πιθανοφάνειας μοντέλου και μιας μεγιστοποιημένης λογαριθμικής πιθανοφάνειας
Leblanc & Crowley (1993)	Ελεγχοσυνάρτηση logrank με την εισαγωγή ενός μέτρου πολυπλοκότητας διαχωρισμού για ουσιαστικό «κλάδεμα» του δέντρου επιβίωσης
Loh (1991) και Ahn & Loh (1994)	Δύο κριτήρια διαίρεσης βασισμένα στα κατάλοιπα από προσαρμογή μονοδιάστατου μοντέλου Cox κάθε φορά και τη διερεύνηση της επεξηγηματικής μεταβλητής με λιγότερο τυχαίο μοτίβο αξόνων
Zhang (1995)	Δύο μέτρα πρόσμιξης ένα για τους παρατηρούμενους χρόνους και ένα για τους λογοκριμένους
Keles & Segal (2002)	Χρήση συναρτήσεων διαίρεσης logrank και συναρτήσεων αθροίσματος τετραγώνων καταλοίπων martingale που επαναϋπολογίζονται σε κάθε κόμβο
Molinaro et al. (2004)	Παρατηρούμενη συνάρτηση απώλειας με λογοκρισία σταθμίζοντας μια πλήρη συνάρτηση απώλειας χωρίς λογοκρισία, γνωστή ως IPCW (Inverse Probability of Censoring Weights)
Jin et al. (2004)	Κανόνας διαχωρισμού που βασίζεται στη διακύμανση των χρόνων επιβίωσης (χρήση ενός περιορισμένου χρονικού ορίου για τον υπολογισμό της διακύμανσης λόγω λογοκριμένων παρατηρήσεων)
Cho & Hong (2008)	Συνάρτηση απώλειας L_1 για να κατασκευάσουν ένα διάμεσο δέντρο επιβίωσης
Zhu & Kosorok (2012)	Αναδρομική παλινδρόμηση δέντρου επιβίωσης (RIST): μη παραμετρική διαδικασία παλινδρόμησης που χρησιμοποιεί μια νέα αναδρομική προσέγγιση απόδοσης (imputation) σε συνδυασμό με εξαιρετικά τυχαιοποιημένα δέντρα
Steingrimsson et al. (2016)	Διπλά ισχυρές επεκτάσεις των εκτιμητών απώλειας IPCW που παρουσιάζονται στην εργασία των Molinaro et al. (2004), γνωστές ως L_{DR}
Steingrimsson et al. (2019)	Κατασκευή αμερόληπτων ως προς τη λογοκρισία συναρτήσεων απώλειας για την επέκταση των αλγορίθμων CART (Classification and Regression Trees)
Sun et al. (2020)	Χρήση κριτηρίων ΔICON και ICON για τη δημιουργία των δέντρων επιβίωσης με καθοδήγηση από τις καμπύλες ROC και μεταβλητές είτε εξαρτώμενες από το χρόνο είτε ανεξάρτητες.

Πίνακας 3.1. **Χρονολογική εξέλιξη των δέντρων επιβίωσης (1985 – 2020)**

Μια ριζικά διαφορετική προσέγγιση στο πεδίο των δενδρικών δομών αποτελεί η εργασία των Hothorn et al. (2006), μέσω της οποίας επιχειρήθηκε η επίλυση δύο

θεμελιωδών προβλημάτων που αντιμετωπίζουν οι κλασικές ευρετικές τεχνικές τύπου CART: την υπερπροσαρμογή και τη μεροληψία επιλογής μεταβλητών (selection bias) υπέρ εκείνων που παρουσιάζουν πολλές πιθανές διασπάσεις ή ελλείπουσες τιμές. Ενώ το πρόβλημα της υπερπροσαρμογής αντιμετωπίζεται εν μέρει με τεχνικές κλαδέματος (pruning), η μεροληψία επιλογής δυσχεραίνει σημαντικά την ερμηνευσιμότητα και κατ' επέκταση την προγνωστική ισχύ του υποδείγματος. Για την άρση αυτών των περιορισμών, οι Hothorn et al. (2006) πρότειναν τα Δέντρα Υπό Όρους Συμπερασμού (Conditional Inference Trees – CI Trees), τα οποία υιοθετούν ένα υπολογιστικά αποδοτικό στατιστικό πλαίσιο. Στην αρχιτεκτονική αυτή, η επιλογή των μεταβλητών διαχωρισμού πραγματοποιείται με απόλυτα αμερόληπτο τρόπο, καθώς βασίζεται σε αυστηρούς ελέγχους στατιστικής σημαντικότητας (permutation tests) αντί για μεγιστοποίηση ενός κριτηρίου καθαρότητας. Πιο συγκεκριμένα, στην περίπτωση που όλες οι επεξηγηματικές μεταβλητές είναι ανεξάρτητες από τη μεταβλητή απόκρισης, η ενσωμάτωση ενός αυστηρού κριτηρίου διακοπής (stopping criterion) βάσει του επιπέδου σημαντικότητας α καταφέρνει να μειώνει αισθητά το ποσοστό των λανθασμένων διασπάσεων στον εκάστοτε κόμβο. Αντίθετα, στην περίπτωση όπου κάποια συμμεταβλητή εμφανίζει πραγματική στατιστική συσχέτιση με την απόκριση, επιλέγεται με αυξημένη συχνότητα σε σχέση με την κλασική, εξαντλητική διαδικασία αναζήτησης, διασφαλίζοντας τη σταθερότητα της δομής.

3.2.1 Βέλτιστα Δέντρα Επιβίωσης

Μια σύγχρονη και καινοτόμος τεχνική στην κατηγορία αυτή είναι τα Βέλτιστα Δέντρα Επιβίωσης (Optimal Survival Trees – OST). Τα OST προτάθηκαν αρχικά από τους Bertsimas et al. (2022) με στόχο την κατασκευή καθολικά βελτιστοποιημένων (globally optimized) μοντέλων. Η προσέγγιση αυτή επιτυγχάνεται μέσω της χρήσης μεικτής ακέραιας βελτιστοποίησης (mixed – integer optimization – MIO) και τοπικής αναζήτησης, καθιστώντας τη διαδικασία ιδιαίτερα κατάλληλη για ανάλυση δεδομένων επιβίωσης υψηλής διάστασης. Η κεντρική ιδέα των Βέλτιστων Δέντρων (Optimal Trees) αναπτύχθηκε αρχικά από τους Bertsimas & Dunn (2019) και Dunn (2018). Βασικός στόχος των ερευνητών ήταν η αντιμετώπιση της «απληστίας» (greediness) που χαρακτηρίζει τον παραδοσιακό αλγόριθμο CART. Στον CART, η βελτιστοποίηση της διαίρεσης εκτελείται τοπικά σε κάθε κόμβο, αγνοώντας την πιθανή επίδραση των επόμενων διαχωρισμών, γεγονός που συχνά οδηγεί σε υποβέλτιστα δέντρα και προβλήματα γενικευσιμότητας. Αντίθετα, η μεθοδολογία των OST επιτρέπει στο μοντέλο να ανταγωνίζεται σε προγνωστική ισχύ συνδυαστικές μεθόδους, όπως τυχαία δάση (Random Forests) και τα ενισχυμένα δέντρα (Gradient Boosted Trees) (βλ. Ενότητα 3.3 και 3.4) διατηρώντας παράλληλα την πλήρη ερμηνευσιμότητα ενός μεμονωμένου τελικού δέντρου.

Οι Bertsimas et al. (2022) επέκτειναν αυτή τη φιλοσοφία στο πεδίο της Ανάλυσης Επιβίωσης, όπου η ερμηνευσιμότητα είναι κρίσιμη για τη λήψη κλινικών αποφάσεων από τους ιατρούς. Η κύρια πρόκληση στην εν λόγω διαδικασία είναι η διαχείριση των λογοκριμένων παρατηρήσεων. Ο κανόνας διαχωρισμού που εφαρμόζεται βασίζεται στον ακόλουθο τύπο:

$$P(S_i \leq t) = 1 - e^{-\theta_i \Lambda(t)}$$

με $\Lambda(t)$: αναφορική c.h.f. στο χρόνο t και οι συντελεστές θ_i: προσαρμογές στη $\Lambda(t)$.

Η εκτίμηση της $\Lambda(t)$ βασίζεται στον εκτιμητή Nelson – Aalen και υπολογίζεται ως εξής:

$$\hat{\Lambda}(t) = \sum_{i:t_i \leq t} \frac{\delta_i}{\sum_{j:t_j \geq t_i} 1}$$

ενώ οι συντελεστές θ_k εκτιμώνται με τη μεγιστοποίηση της συνάρτησης πιθανοφάνειας δείγματος εντός του κόμβου (within – leaf sample likelihood):

$$L = \prod_{i=1}^{n} \left(\theta_i \frac{d}{dt} \Lambda(t_i) \right)^{\delta_i} e^{-\theta_i \Lambda(t_i)}$$

Η κλειστή μορφή του υπολογισμού τους δίνεται από τη σχέση:

$$\hat{\theta}_k = \frac{\sum_i \delta_i I(T_i = k)}{\sum_i \hat{\Lambda}(t_i) I(T_i = k)}$$

Για την αξιολόγηση των διαχωρισμών, κατασκευάζονται n πλήρως κορεσμένα δέντρα επιβίωσης για κάθε παρατήρηση. Η εκτίμηση των $\boldsymbol{\theta^{sat}}$ των συγκεκριμένων δέντρων υπολογίζεται ως εξής:

$$\hat{\theta}_i^{sat} = \frac{\delta_i}{\hat{\Lambda}(t_i)}, i = 1, \dots, n$$

Για τη βελτιστοποίηση του δέντρου γίνεται ελαχιστοποίηση της συνολικής συνάρτησης σφάλματος:

$$error(T, D) = \sum_{k \in \boldsymbol{leaves}(T)} error_k(D) \qquad \mu\varepsilon$$

$$error_k = \sum_{i:T(i)=k} \left(\delta_i \log\left(\frac{\delta_i}{\hat{\Lambda}(t_i)} \right) - \delta_i \log(\hat{\theta}_k) - \delta_i + \hat{\theta}_k \hat{\Lambda}(t_i) \right)$$

Βάσει των παραπάνω, ο αλγόριθμος του Βέλτιστου Δέντρου εφαρμόζεται αντικαθιστώντας την τιμή αυτή στη συνολική συνάρτηση απώλειας:

$$\min_T \ error(T, D) + a \cdot complexity(T)$$

όπου T είναι το υπό βελτιστοποίηση δέντρο απόφασης, D είναι το σύνολο δεδομένων εκπαίδευσης (training set), η συνάρτηση $error(T, D)$ μετρά την ποιότητα προσαρμογής του δέντρου στα δεδομένα D, ενώ η $complexity(T)$ συνάρτηση – παράμετρος πολυπλοκότητας του δέντρου που εξισορροπεί την ακρίβεια του δέντρου με το μέγεθος του. Τέλος, η εφαρμογή μιας εσωτερικής διαδικασίας διασταυρούμενης

επικύρωσης CV κρίνεται απαραίτητη για την ορθή αξιολόγηση και επιλογή του τελικού μοντέλου.

Στα βασικά πλεονεκτήματα της συγκεκριμένης τεχνικής συγκαταλέγεται το γεγονός ότι, καθώς αυξάνεται ο αριθμός των παρατηρήσεων, γενικευσιμότητα του μοντέλου βελτιώνεται σταθερά. Αντίθετα, οι κλασικοί ευρετικοί αλγόριθμοι τείνουν να είναι ιδιαίτερα επιρρεπείς στην υπερπροσαρμογή κατά τη διαχείριση μεγάλων συνόλων δεδομένων. Τέλος, οι Bertsimas et al. (2022) έχουν αναπτύξει το εξειδικευμένο λογισμικό "Interpretable AI" μέσω του οποίου παρέχεται η δυνατότητα πρακτικής εφαρμογής όλων των μεθοδολογιών βελτιστοποίησης που έχει προτείνει η συγκεκριμένη ερευνητική ομάδα.

3.2.2 Βέλτιστα Αραιά Δέντρα Επιβίωσης

Ως εξέλιξη των Βέλτιστων Δέντρων Επιβίωσης (OST) που προτάθηκαν από τους Bertsimas et al. (2022), αναπτύχθηκε η εργασία των Zhang et al. (2024), η οποία εισάγει τα Βέλτιστα Αραιά Δέντρα Επιβίωσης (Optimal Sparse Survival Tree – OSST). Οι Zhang et al. (2024) επισημαίνουν ορισμένους σημαντικούς περιορισμούς της προσέγγισης OST:

1) Η αρχιτεκτονική OST βασίζεται στην υπόθεση ότι η αναλογία των συναρτήσεων κινδύνου για οποιαδήποτε δύο άτομα πρέπει να παραμείνει σταθερή με την πάροδο του χρόνου (υπόθεση αναλογικότητας),
2) Κατά τη διαδικασία αναζήτησης του βέλτιστου, οι ευρετικοί αλγόριθμοι ενδέχεται να εγκλωβιστούν σε τοπικά ακρότατα (local optima).
3) Η εφαρμογή της μεθοδολογίας OST βασίζεται σε ιδιόκτητο κώδικα κλειστού λογισμικού, ο οποίος παρουσιάζει περιορισμούς κλιμάκωσης σε μεγάλα σύνολα δεδομένων.

Για την αντιμετώπιση αυτών των ζητημάτων, οι Zhang et al. (2024) πρότειναν έναν αλγόριθμο ανοιχτού κώδικα βασισμένο σε οριοθετημένο δυναμικό προγραμματισμό (bounded dynamic programming). Ο αλγόριθμος χρησιμοποιεί ως συνάρτηση απώλειας το Ολοκληρωμένο Σκορ Brier (Integrated Brier Score) των Graf et al. (1999) για τη μείωση του αριθμού των κόμβων και του υπολογιστικού κόστους.

Για την τυπική διατύπωση της τεχνικής, ορίζεται το σύνολο δεδομένων εκπαίδευσης $(\boldsymbol{X}, \boldsymbol{c}, \boldsymbol{y})$ ως $\{(\boldsymbol{x}_i, c_i, y_i)\}_{i=1}^{N}$, όπου $c_i = \{0,1\}$ είναι ο δείκτης λογοκρισίας, με $c_i = 1$ είναι ο παρατηρούμενος πλήρης χρόνος ενώ $c_i = 0$ διαφορετικά και $y_i \in \mathbb{R}$ είναι ο χρόνος της τελευταίας παρατήρησης του δείγματος i και $\boldsymbol{x}_i = \{0,1\}^M$ αποτελεί ένα δυαδικό διάνυσμα χαρακτηριστικών. Σημειώνεται ότι για την εφαρμογή του αλγόριθμου, οι συνεχείς μεταβλητές πρέπει να μετατραπούν σε ένα σύνολο δυαδικών μεταβλητών, χρησιμοποιώντας είτε όλες τις πιθανές διασπάσεις (thresholds) είτε ένα υποσύνολο αυτών που επιλέγονται μέσω ενός μοντέλου Black Box (McTavish, et al., 2022). Η εμπειρική απώλεια (loss) ενός δέντρου t στο σύνολο δεδομένων εκπαίδευσης ορίζεται ως το Integrated Brier Score:

$$\mathcal{L}(t, \boldsymbol{X}, \boldsymbol{c}, \boldsymbol{y}) \coloneqq \frac{1}{y_{max}} \int_0^{y_{max}} BS(y)dy\,, \qquad \text{όπου}$$

- $y_{max} = \max\{y_i\}_{i=1}^N$ είναι το τελευταίο χρονικό σημείο όλων των παρατηρούμενων δειγμάτων
- $BS(y) = \frac{1}{N}\sum_{i=1}^N \frac{\left(\hat{S}_{\boldsymbol{X_i}}(y)-0\right)^2}{\hat{G}(y_i)} \cdot I(y_i \le y, c_i = 1) + \frac{\left(\hat{S}_{\boldsymbol{X_i}}(y)-1\right)^2}{\hat{G}(y)} \cdot I(y_i > y)$: είναι το Brier Score του δέντρου t στο χρόνο y και εκφράζεται ως το μέσο τετραγωνικό σφάλμα μεταξύ των δεδομένων και της προβλεπόμενης s.f. $\hat{S}_{\boldsymbol{X_i}}(y)$ του δέντρου t, σταθμισμένο με την Αντίστροφη Πιθανότητα Λογοκρισίας (IPCW)
- $\hat{S}_{\boldsymbol{X_i}}(\cdot)$: μη – παραμετρική εκτίμηση Kaplan – Meier της s.f. στον αντίστοιχο κόμβο του δέντρου με χαρακτηριστικά $\boldsymbol{X_i}$
- $\hat{G}(\cdot)$: εκτίμηση Kaplan – Meier για την κατανομή λογοκρισίας του $\boldsymbol{c}$, το οποίο θεωρείται ανεξάρτητο από κάθε επεξηγηματική μεταβλητή (Molinaro et al. (2004)).

Με στόχο την εύρεση ενός δέντρου που εξισορροπεί την ακρίβεια και την απλότητα (sparsity), ορίζεται η τελική αντικειμενική συνάρτηση $R(t, \boldsymbol{X}, \boldsymbol{y})$ του δέντρου t ως εξής:

$$R(t, \boldsymbol{X}, \boldsymbol{y}) \coloneqq \mathcal{L}(t, \boldsymbol{X}, \boldsymbol{c}, \boldsymbol{y}) + \lambda \cdot complexity(t), \qquad με\ depth(t) \le d$$

όπου d: είναι ο αυστηρός περιορισμός στο βάθος του δέντρου, ο οποίος όσο μικρότερη είναι η τιμή, τόσο ταχύτερη είναι η σύγκλιση του αλγορίθμου και τόσο μικρότερο (πιο ερμηνεύσιμο) το τελικό δέντρο.

Στα βασικά πλεονεκτήματα της συγκεκριμένης μεθοδολογίας συγκαταλέγονται η εξαιρετική δυνατότητα γενικευσιμότητας των αποτελεσμάτων (γενικότερη αποφυγή του overfitting λόγω της ποινής λ) καθώς και η ομοιόμορφη υπολογιστική κλιμάκωση σε μεγάλα σύνολα δεδομένων. Ως μελλοντικές κατευθύνσεις για βελτίωση, οι Zhang et al. (2024) προτείνουν την επέκταση της αρχιτεκτονικής και σε άλλα μέτρα απόδοση, καθώς και την ανάπτυξη ακόμη ταχύτερων αλγορίθμων που διατηρούν αναλλοίωτη τη μαθηματική ακρίβεια του παγκόσμιου βέλτιστου.

3.3 Τυχαία Δάση Επιβίωσης

Στην προηγούμενη ενότητα παρουσιάστηκε η θεμελιώδης αρχιτεκτονική των Δέντρων Επιβίωσης, καθώς και ποικίλες μεθοδολογικές τροποποιήσεις αυτών. Το κυριότερο πλεονέκτημα των δενδρικών δομών έγκειται στην υψηλή ερμηνευσιμότητα που προσφέρουν στους αναλυτές. Ωστόσο, τα απλά δέντρα επιβίωσης συχνά στερούνται γενικευσιμότητας λόγω του φαινομένου της υπερπροσαρμογής (overfitting), το οποίο εμφανίζεται συχνότερα σε περιβάλλοντα δεδομένων υψηλής διάστασης ($p \gg n$). Επιπλέον, η στοχαστική προσέγγιση ενός μεμονωμένου δέντρου ενδέχεται να οδηγήσει είτε σε υπερβολικά εκτενείς δομές με δυσκολία ερμηνείας, είτε σε υπερβολικά απλούστερες δομές που παραλείπουν κρίσιμη πληροφορία. Το

σημαντικότερο, ωστόσο, μειονέκτημα των μεμονωμένων δέντρων είναι η στατιστική τους αστάθεια όσον αφορά τις προκύπτουσες προβλέψεις. Συγκεκριμένα, ακόμη και μικρές μεταβολές στο δείγμα εκπαίδευσης μπορούν να οδηγήσουν σε σημαντικές αλλαγές στη δομή του δέντρου και, κατ' επέκταση, στη προβλεπτική συνάρτηση.

Μια αποτελεσματική προσέγγιση για την άρση των περιορισμών προτάθηκε από τον Breiman (2001) με την εισαγωγή των Τυχαίων Δασών (Random Forest – RF). Η κεντρική ιδέα των τυχαίων δασών βασίζεται στην ανάπτυξη ενός προκαθορισμένου αριθμού δέντρων B, η βέλτιστη τιμή του οποίου ορίζεται είτε εμπειρικά είτε μέσω τεχνικών διασταυρούμενης επικύρωσης CV. Κάθε μεμονωμένο δέντρο εκπαιδεύεται σε ένα τυχαίο δείγμα αναδειγματοληψίας με αντικατάσταση (bootstrap sample) από τα αρχικά δεδομένα. Παράλληλα, σε κάθε κόμβο του δέντρου, η αναζήτηση του βέλτιστου διαχωρισμού δεν πραγματοποιείται στο σύνολο των μεταβλητών, αλλά σε ένα τυχαίο επιλεγμένο υποσύνολο αυτών. Το τελικό αποτέλεσμα του υποδείγματος προκύπτει από τη συνάθροιση (aggregation) και τη λήψη του μέσου όρου των προβλέψεων όλων των δέντρων. Μια εναλλακτική διαδικασία αποτελεί η διαδικασία του bagging (bootstrap aggregation), η οποία διαφοροποιείται ως προς το δεύτερο στάδιο τυχαιοποίησης. Ενώ διατηρείται η λήψη bootstrap δειγμάτων για κάθε δέντρο, οι υποψήφιες μεταβλητές για διαχωρισμό σε κάθε κόμβο περιλαμβάνουν το σύνολο και όχι ένα τυχαίο υποσύνολο των επεξηγηματικών μεταβλητών. Ο Breiman (2001) απέδειξε ότι η μάθηση συνόλου (ensemble learning) βελτιώνεται δραστικά μέσω της εισαγωγής στοιχείων τυχαιοποίησης στη βασική διαδικασία εκπαίδευσης, επιτυγχάνοντας σημαντική μείωση της διακύμανσης χωρίς παράλληλη αύξηση της μεροληψίας (bias). (Bou – Hamad et al. (2011))

Βασιζόμενοι σε αυτό το πλαίσιο, οι Ishwaran et al. (2008) επέκτειναν την ανωτέρω αρχιτεκτονική εισάγοντας τα Τυχαία Δάση Επιβίωσης (Random Survival Forest – RSF) για τη διαχείριση δεδομένων με δεξιά λογοκρισία. Η μεθοδολογία RSF ενσωματώνει καινοτόμους μηχανισμούς, όπως τη μέτρηση της σπουδαιότητας μεταβλητών (Variable Importance – VIMP) και έναν προσαρμοστικό αλγόριθμο καταλογισμού ελλειπουσών τιμών (adaptive tree imputation), ο οποίος επιτρέπει τη διαχείριση εκτενούς έλλειψης δεδομένων παράγοντας σχεδόν αμερόληπτες εκτιμήσεις του σφάλματος πρόβλεψης. Στα RSF, η εκτίμηση της c.h.f. στους τερματικούς κόμβους κάθε δέντρου πραγματοποιείται μέσω του μη παραμετρικού εκτιμητή Nelson – Aalen, ενώ η συνολική πρόβλεψη του δάσους (ensemble c.h.f.) προκύπτει ως μέσος όρος των επιμέρους εκτιμήσεων. Στη συνέχεια, οι Ishwaran et al. (2011) εξειδίκευσαν τα RSF για δεδομένα υψηλής διάστασης, εισάγοντας μια αδιάστατη στατιστική τάξης (dimensionless order statistic) για την επιλογή μεταβλητών, η οποία ονομάζεται ελάχιστο βάθος (minimal depth) (Ishwaran et al. (2010)). Το μέτρο αυτό αξιολογεί την προγνωστική ισχύ μιας μεταβλητής εντός του δάσους, συμπληρώνοντας τη μετρική VIMP, η οποία υπολογίζει τη μεταβολή στο σφάλμα πρόβλεψης του συνόλου όταν οι τιμές μιας μεταβλητής διαταράσσονται τυχαία. Γεωμετρικά, το ελάχιστο βάθος ορίζεται ως η μικρότερη απόσταση (αριθμός ακμών) από την κορυφή (ρίζα) του δέντρου έως τον αρχικό κόμβο του πλησιέστερου μέγιστου υποδέντρου (maximal subtree) που αντιστοιχεί στη συγκεκριμένη

μεταβλητή. Ως μέγιστο υποδέντρο ορίζεται το μεγαλύτερο δυνατό υποδέντρο του οποίου ο πρωτεύων (ριζικός) διαχωρισμός πραγματοποιείται βάσει της εξεταζόμενης μεταβλητής. Με βάση την ιδιότητα του ελάχιστου βάθους, οι ερευνητές πρότειναν κατάλληλες τροποποιήσεις στα RSF, καθιστώντας τα ένα ισχυρό εργαλείο ταυτόχρονης πρόβλεψης και επιλογής χαρακτηριστικών (feature screening). Ωστόσο, ένα βασικό μειονέκτημα που χαρακτήριζε διαχρονικά τη μετρική VIMP στο πλαίσιο των κλασικών αρχιτεκτονικών RF και RSF ήταν η εγγενής δυσκολία ακριβούς εκτίμησης της διακύμανσής της. Για την επίλυση αυτού του περιορισμού, οι Ishwaran & Lu (2019) μια καινοτόμο προσέγγιση βασισμένη σε τεχνικές υποδειγματοληψίας. Συγκεκριμένα, μέσω της ενσωμάτωσης του jackknife με διαγραφή d παρατηρήσεων (delete – d jackknife estimator), οι ερευνητές πέτυχαν να εκτιμήσουν με ακρίβεια τη διακύμανση του μέτρου σπουδαιότητας και, κατ' επέκταση, να κατασκευάσουν τα αντίστοιχα διαστήματα εμπιστοσύνης. Με τον τρόπο αυτό, η αξιολόγηση και η επιλογή των μεταβλητών βάσει του VIMP παύει να είναι μια αποκλειστικά ευρετική διαδικασία και αποκτά μια συμπαγή στατιστική βάση, επιτρέποντας στον αναλυτή να αιτιολογεί κάθε απόφαση επιλογής χαρακτηριστικών με σαφώς ορισμένο περιθώριο σφάλματος.

Μία εναλλακτική και ιδιαίτερα αποδοτική προσέγγιση στο πεδίο των συνόλων επιβίωσης αποτελεί η εργασία των Zhou et al. (2015) οι οποίοι ανέπτυξαν τα Δάση Επιβίωσης Περιστροφής (Rotation Survival Forest – RotSF) για δεδομένα με δεξιά λογοκρισία. Η τεχνική RotSF εκτελείται σε δύο στάδια:

1. Αρχικά, λαμβάνεται ένα bootstrap δείγμα του συνόλου δεδομένων εκπαίδευσης για κάθε δέντρο του δάσους
2. Σε δεύτερη φάση, το σύνολο των επεξηγηματικών μεταβλητών διαιρείται τυχαία σε έναν αριθμό ασύνδετων υποσυνόλων. Σε κάθε υποσύνολο εφαρμόζεται Ανάλυση Κύριων Συνιστωσών (Principal Component Analysis – PCA) με σκοπό την εξαγωγή ιδιοδιανυσμάτων που συγκροτούν τον πίνακα περιστροφής (rotation matrix).

Ο γραμμικός πολλαπλασιασμός του αρχικού πίνακα δεδομένων με τον πίνακα περιστροφής οδηγεί σε έναν νέο, μετασχηματισμένο χώρο χαρακτηριστικών, επί του οποίου εκτελείται ο αλγόριθμος ανάπτυξης των δέντρων επιβίωσης. Στα κυριότερα μειονεκτήματα της μεθόδου RotSF συγκαταλέγεται το υψηλό υπολογιστικό κόστος που προκύπτει από τις επαναλαμβανόμενες εφαρμογές της PCA, αν και η δυνατότητα παράλληλης επεξεργασίας των δέντρων καθιστά την προσέγγιση εφικτή σε μεγάλα σύνολα δεδομένων. Επιπλέον, ο προσδιορισμός της παραμέτρου που ελέγχει το μέγεθος των υποσυνόλων των μεταβλητών στερείται ενός τυποποιημένου αναλυτικού κανόνα. Προκειμένου να αμβλύνουν το υπολογιστικό κόστος της RotSF σε περιβάλλοντα εξαιρετικά υψηλής διάστασης, οι Zhou et al. (2015) εξέλιξαν τη μέθοδο παρουσιάζοντας τα Τυχαία Δάση Επιβίωσης Περιστροφής (Random Rotation Survival Forest – RRotSF). Η αρχιτεκτονική RRotSF αντιμετωπίζει τις απαιτητικές πράξεις της PCA συνδυάζοντας τις αρχές του δάσους περιστροφής, το bagging και τον τυχαίο υποχώρο (random subspace). Ο όρος «τυχαίος υποχώρος» αναφέρεται στη διαδικασία τυχαίας επιλογής μιας υποομάδας χαμηλότερης διάστασης από το αρχικό

πλήθος των επεξηγηματικών μεταβλητών πριν την εφαρμογή του μετασχηματισμού περιστροφής, περιορίζοντας δραστικά τις υπολογιστικές απαιτήσεις.

Μια ακόμη σημαντική συνεισφορά προτάθηκε από τους Steingrimsson et al. (2019) οι οποίοι επέκτειναν τη θεωρία CUT (Censoring Unbiased Transformations), που είχε εισαχθεί από τους Fan & Gijbels (1994), στην περίπτωση των δέντρων παλινδρόμησης και των μεθόδων μάθησης συνόλου. Συγκεκριμένα ανέπτυξαν τις μεθοδολογίες CURT και CURE (Censored Unbiased Regression Trees and Ensembles), μέσω των οποίων οι συναρτήσεις απώλειας για πλήρη δεδομένα αντικαθίστανται από κατάλληλες συναρτήσεις απώλειας, επιτρέποντας την επέκταση των αλγορίθμων CART και RF σε δεδομένα με λογοκριμένες αποκρίσεις. Οι συγγραφείς εξετάζουν, μεταξύ άλλων, δύο προσεγγίσεις για την κατασκευή των απαιτούμενων συναρτήσεων απώλειας, βασισμένες στη μεθοδολογία Buckley – James και σε μια «διπλά εύρωστη» (doubly robust) προσέγγιση. Για την ειδική περίπτωση της συνάρτησης απώλειας L_2, δείχνουν επιπλέον ότι μπορεί να χρησιμοποιηθεί μια συγκεκριμένη μορφή imputation στην απόκριση, επιτρέποντας την υλοποίηση των προτεινόμενων αλγόριθμων με τυπικό λογισμικό που έχει σχεδιαστεί για μη λογοκριμένες αποκρίσεις. Οι προτεινόμενες μέθοδοι είχαν καλή προβλεπτική απόδοση τόσο σε προσομοιωμένα δεδομένα όσο και σε πραγματικά σύνολα δεδομένων, σε σύγκριση με υφιστάμενες μεθόδους μάθησης συνόλου.

Τέλος, μια εξαιρετικά σημαντική εναλλακτική προσέγγιση των Τυχαίων Δασών, η οποία αποτελεί και άμεση επέκταση των CI Trees, εντοπίζεται στις κομβικές εργασίες των Strobl et al. (2007) και Strobl et al. (2008), οι οποίες εισήγαγαν τα Δάση υπό Όρους Συμπερασμού (Conditional Inference Forests – CI Forests). Συγκεκριμένα, στόχος της πρώτης μελέτης (Strobl et al. (2007)) ήταν η αντιμετώπιση της μεροληψίας που προέκυπτε εξαιτίας της ποικιλομορφίας (διαφορετική κλίμακα, τύπος ή πλήθος κατηγοριών) των επεξηγηματικών μεταβλητών. Μέχρι τότε, τα αποτελέσματα των μέτρων σπουδαιότητας (VIMP) στα κλασικά τυχαία δάση δεν ήταν αξιόπιστα, καθώς η αρχιτεκτονική τους είχε σχεδιαστεί πρωτίστως για τη βελτιστοποίηση της πρόβλεψης και όχι για την αμερόληπτη επιλογή μεταβλητών (variable selection). Οι ερευνητές ανέδειξαν ότι η κύρια αιτία αυτής της τεχνικής μεροληψίας εντοπιζόταν στη δειγματοληψία bootstrap. Για το σκοπό αυτό, πρότειναν μια αμερόληπτη διαδικασία επιλογής μεταβλητών στα μεμονωμένα δέντρα μέσω υποδειγματοληψίας χωρίς επανάθεση, αναπτύσσοντας έναν αλγόριθμο υπολογιστικά ταχύτερο και στατιστικά ακριβέστερο σε σχέση με τα παραδοσιακά RF. Στη δεύτερη εργασία τους, οι Strobl et al. (2008) εισήγαγαν νέα διαδικασία μετάθεσης (conditional permutation scheme) για τον υπολογισμό της σπουδαιότητας των μεταβλητών. Μέσω αυτού, δημιούργησαν μια στατιστική διαδικασία φιλτραρίσματος η οποία απομονώνει τις επιδράσεις αλληλεξάρτησης μεταξύ των επεξηγηματικών μεταβλητών. Αντίθετα με την αρχική, οριακή προσέγγιση των κλασικών δασών, η οποία τείνει να υπερεκτιμά τη σπουδαιότητα μεταβλητών λόγω έντονης πολυσυγγραμμικότητας, το υπό όρους σχήμα των Strobl et al. (2008) αντανακλά την πραγματική, αυτόνομη επίδραση της εκάστοτε επεξηγηματικής μεταβλητής στη μελετώμενη μεταβλητή απόκρισης. Τα επόμενα έτη, η επιστημονική έρευνα επικεντρώθηκε στη

βελτιστοποίηση του αλγορίθμου, με στόχο την ανάπτυξη υπολογιστικά πιο αποδοτικών και ταχύτερων υλοποιήσεων, διατηρώντας παράλληλα ανέπαφη την κεντρική φιλοσοφία της μεθόδου. Στη σύγχρονη βιβλιογραφία, οι CI Forests και οι RSF αλγόριθμοι αποτελούν τα δύο βασικά σημεία αναφοράς, έχοντας αποτελέσει το αντικείμενο πολυάριθμων συγκριτικών μελετών σε δεδομένα υψηλής διάστασης.

3.3.1 Πλάγιο Τυχαίο Δάσος Επιβίωσης

Μια εναλλακτική και ταυτόχρονα ιδιαίτερα ισχυρή τεχνική ως προς την προγνωστική της αξία σε σχέση με τα RSF αποτελεί η εργασία των Jaeger et al. (2019), οι οποίοι ανέπτυξαν τη μεθοδολογία των Πλάγιων Τυχαίων Δασών Επιβίωσης (Oblique Random Survival Forests – ORSF). Οι Jaeger et al. (2019) βάσισαν τη διατύπωση της ιδέας τους στη χρήση κανονικοποιημένων μοντέλων αναλογικού κινδύνου του Cox, τα οποία ενσωματώνονται στους μη τερματικούς κόμβους των δέντρων επιβίωσης. Υπό την προϋπόθεση ότι σε έναν ενδιάμεσο κόμβο υπάρχουν K παρατηρήσεις, το ενσωματωμένο μοντέλο Cox λαμβάνει τη μορφή:

$$h_k(t) = h_0(t)e^{\boldsymbol{x}_k^T\boldsymbol{\beta}}$$

όπου $k = 1, \dots, K$, $h_k(t)$ είναι η h.f. της παρατήρησης k στο χρόνο t, $h_0(t)$ η αναφορική h.f. και $\boldsymbol{\beta}$ είναι ένα διάνυσμα των $mtry \leq p$ συντελεστών με $mtry$ να είναι ο αριθμός των τυχαία επιλεγμένων μεταβλητών πρόβλεψης για το συγκεκριμένο κόμβο και p ο αριθμός όλων των διαθέσιμων επεξηγηματικών μεταβλητών. Η εκτίμηση των $\boldsymbol{\beta}$ προκύπτει από μεγιστοποίηση της ποινικοποιημένης συνάρτησης μερικής πιθανοφάνειας:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{m} \frac{e^{\boldsymbol{x}_{j(i)}^T\boldsymbol{\beta}}}{\sum_{j\in R_i} e^{\boldsymbol{x}_j^T\boldsymbol{\beta}}}$$

όπου R_j είναι το σύνολο των παρατηρήσεων που βρίσκονται σε κίνδυνο τη χρονική στιγμή t_i, m είναι ο αριθμός των μοναδικών χρόνων που συνέβη η υπό μελέτη έκβαση (υποθέτοντας απουσία δεσμών) και $j(i)$ είναι ο δείκτης της παρατήρησης στην οποία εκδηλώθηκε το υπό μελέτη συμβάν τη χρονική στιγμή t_i. Ο περιορισμός της παραπάνω συνάρτησης ορίζεται η ποινή elastic net (Zou & Hastie, 2005):

$$\alpha \sum_{i=1}^{p} |\beta_i| + (1-\alpha) \sum_{i=1}^{p} \beta_i^2 \leq s, \qquad s \geq 0$$

με τη σταθερά s να ελέγχει το συνολικό βαθμό συρρίκνωσης και να έχει αντιστοιχία 1 – 1 με την υπερπαράμετρο $\lambda \geq 0$. Ως γνωστόν, για $\alpha = 1$ και $\alpha = 0$ προκύπτουν οι ποινές Lasso και Ridge αντίστοιχα. Παραπάνω σχόλια και παρατηρήσεις για τις συγκεκριμένες ποινές παρουσιάζονται εκτενέστερα στην ενότητα 2.4.

Στη συνέχεια, οι Jaeger et al. (2019) εντάσσουν στην αρχιτεκτονική των δέντρων την έννοια των γραμμικών συνδυασμών των μεταβλητών εισόδου (Linear combinations of input variables – LCIVs). Σε κάθε κόμβο, υπάρχουν K υποψήφιες μεταβλητές για διαχωρισμό του κόμβου, με $\boldsymbol{x}^T\widehat{\boldsymbol{\beta}}$ όπου $\boldsymbol{x}^T = (\boldsymbol{x}_1^T, \dots, \boldsymbol{x}_K^T)$ να είναι ένας

στοιβαγμένος πίνακας διανυσμάτων εισόδου, ενώ οι συντελεστές $\widehat{\boldsymbol{\beta}}$ επιλέγονται από έναν αριθμό υποψήφιων λύσεων που δίνονται από το ενσωματωμένο μοντέλο Cox. Ο μέγιστος αριθμός μεταβλητών πρόβλεψης είναι ίσος με το $mtry$ και ο τρόπος εκτίμησης των $\widehat{\boldsymbol{\beta}}$ εξαρτάται άμεσα από την εφαρμογή ή μη διαδικασίας CV και από την τιμή του α. Με την εφαρμογή διαδικασίας 5-fold CV σε κάθε μη τερματικό κόμβο, εκτιμώνται δύο τιμές λ:

- λ_{CV}: η τιμή που μεγιστοποιεί τη CV μερική πιθανοφάνεια
- λ_{SE}: η μέγιστη τιμή του λ εντός ενός τυπικού σφάλματος του λ_{CV}

Η επιλογή της παραμέτρου α επηρεάζει καθοριστικά τις παραπάνω τιμές. Για τιμές του α κοντά στο 1 προκύπτουν εκτιμήσεις τύπου Lasso δηλαδή αρκετοί μηδενικοί συντελεστές άρα και λιγότερες μεταβλητές για επιλογή ως μεταβλητή διαχωρισμού, ενώ για τιμές κοντά στο 0 προκύπτουν εκτιμήσεις τύπου Ridge δηλαδή σχεδόν όλες τις μεταβλητές έχουν μη μηδενικούς συντελεστές. Σε περίπτωση όπου δεν εφαρμόζεται η CV, ακολουθείται μια εναλλακτική διαδικασία κανονικοποίησης για την εκτίμηση των $\lambda_1, \dots, \lambda_{mtry}$ όπου λ_i είναι η μέγιστη τιμή του λ έτσι ώστε το μοντέλο να έχει αποτελεσματικούς βαθμούς ελευθερίας με προαπαιτούμενη τιμή $\alpha < 1$.

Για την αξιολόγηση κάθε υποψήφιας λύσης του διανύσματος $\boldsymbol{\beta}$, ανεξαρτήτως εφαρμογής CV, ακολουθείται η εξής διαδικασία:

1) Υπολογίζεται για κάθε παρατήρηση στον τρέχον κόμβο η ποσότητα $\widehat{\boldsymbol{\eta}} = \boldsymbol{x}^T \widehat{\boldsymbol{\beta}}$.
2) $c_1, \dots, c_{nsplit}$ υποψήφια σημεία διαχωρισμού επιλέγονται τυχαία από τις μοναδικές τιμές του $\widehat{\boldsymbol{\eta}}$.
3) Για $i = 1, \dots, nsplit$ υπολογίζεται η στατιστική log – rank συγκρίνοντας τις καμπύλες επιβίωσης μεταξύ παρατηρήσεων με $\boldsymbol{x}_k^T \widehat{\boldsymbol{\beta}} \leq c_i$ και $\boldsymbol{x}_k^T \widehat{\boldsymbol{\beta}} > c_i$.

Στις περιπτώσεις όπου δεν εφαρμόζεται η CV, εισάγεται μια επιπλέον ποινή στη στατιστική log – rank, $(1 + \gamma \cdot df)$ με γ η παράμετρος συντονισμού και df οι βαθμοί ελευθερίας του μοντέλου που δημιουργεί η υποψήφια LCIV. Η παράμετρος $\gamma > 0$ καθορίζει ουσιαστικά το επίπεδο πολυπλοκότητας και το εύρος του συνόλου επιλογής της τελικής μεταβλητής διαχωρισμού. Βάσει των τελικών τιμών log – rank, ο διαχωρισμός του εκάστοτε κόμβου πραγματοποιείται με κριτήριο τη μέγιστη τιμή της στατιστικής. Στην περίπτωση που καμία από τις υποψήφιες τιμές log – rank δεν υπερβαίνει ένα προκαθορισμένο στατιστικό όριο, ενεργοποιείται κριτήριο πρόωρης διακοπής (early stopping), τερματίζοντας την περαιτέρω ανάπτυξη του συγκεκριμένου υποδέντρου.

Αξίζει να σημειωθεί ότι η τεχνική ORSF χρησιμοποιεί υποδειγματοληψία σε αντίθεση με την κλασική μέθοδο bootstrap με επανάθεση. Σε γενικές γραμμές, οι τεχνικές ORSF και CI Forests παρουσιάζουν σημαντικές μεθοδολογικές συγκλίσεις, καθώς μοιράζονται το ίδιο σχήμα σταθμισμένης συνάθροισης (weighted aggregation scheme). Επιπλέον, και οι δύο τεχνικές εμφανίζουν σχετικά χαμηλό σφάλμα συμφωνίας γεγονός που αποδίδεται στον ενσωματωμένο μηχανισμό πρόωρης

διακοπής (early stopping) στην περίπτωση που δεν διαπιστωθεί ένα ελάχιστο, προκαθορισμένο επίπεδο στατιστικής συσχέτισης. Τέλος, λόγω του πλήθους και της πολυπλοκότητας των μαθηματικών πράξεων που απαιτούνται σε κάθε μη τερματικό κόμβο, η συγκεκριμένη τεχνική συνεπάγεται αντίστοιχα αυξημένο υπολογιστικό χρόνο κατά τη φάση της υλοποίησης.

3.3.2 Τυχαία Δάση για block κλινικών και ωμικών δεδομένων

Η συγκεκριμένη μεθοδολογία προτάθηκε από τους Hornung & Wright (2019) με σκοπό την αποτελεσματική διαχείριση δεδομένων ωμικής (omics data), όπως γονιδιωματικών ή πρωτεωματικών δεδομένων. Τα σύνολα αυτά χαρακτηρίζονται από έναν ιδιαίτερα μεγάλο αριθμό επεξηγηματικών μεταβλητών (γονιδίων) σε σχέση με το μέγεθος του δείγματος, γεγονός που δυσχεραίνει την εφαρμογή των κλασικών τεχνικών μηχανικής μάθησης. Επιπλέον, τα πολυωμικά δεδομένα (multi – omics data) παρουσιάζουν έντονη ετερογένεια, καθώς περιλαμβάνουν διάφορους τύπους μεταβλητών όπως κατηγορικές, συνεχείς και διατεταγμένες μεταβλητές. Η ετερογένεια αυτή εισάγει μεροληψία στους παραδοσιακούς αλγορίθμους, οι οποίοι τείνουν να ευνοούν τις συνεχείς μεταβλητές, παραμελώντας τις υπόλοιπες κατηγορίες. Για την αντιμετώπιση αυτών των περιορισμών, αναπτύχθηκαν πέντε παραλλαγές τυχαίων δασών, προσαρμοσμένες στη δομή block των πολυωμικών συμμεταβλητών. Οι παραλλαγές αυτές διαφοροποιούνται ως προς τον τρόπο επιλογής των υποψήφιων μεταβλητών και των σημείων διαχωρισμού, ενσωματώνοντας τη δομή των blocks.

Οι Hornung & Wright (2019) όρισαν τη δομή των πολυωμικών δεδομένων ως εξής: Έστω ένα σύνολο δεδομένων με n παρατηρήσεις το οποίο αποτελείται από M πίνακες επεξηγηματικών μεταβλητών $\boldsymbol{X}_1, \dots, \boldsymbol{X}_M$ και μεταβλητή απόκρισης $\boldsymbol{y}_1, \dots, \boldsymbol{y}_n$. Ο $m-$οστός πίνακας $\boldsymbol{X}_m$ διάστασης $n \times p_m$ περιέχει, για κάθε παρατήρηση, τις μετρήσεις των p_m μεταβλητών στο $m-$οστό block. Στο πλαίσιο της Ανάλυσης Επιβίωσης, η μεταβλητή απόκρισης εκφράζεται από το διάνυσμα $\boldsymbol{y}_i = \{y_{i,1}, y_{i,2}\}$ όπου το $y_{i,1}$ αντιστοιχεί στο χρόνο επιβίωσης T_i, ενώ το $y_{i,2}$ στη δίτιμη μεταβλητή λογοκρισίας του $i-$οστού ατόμου. Ορίζεται, επίσης, ο πίνακας $\boldsymbol{X} \coloneqq [\boldsymbol{X}_1, \dots, \boldsymbol{X}_M]$ με $p = \sum_{m=1}^{M} p_m$ στήλες.

Οι πέντε εναλλακτικές παραλλαγές που αναπτύχθηκαν ορίζονται ως εξής:

- **VarProb: Πιθανότητες επιλογής μεταβλητών σε συγκεκριμένα block**

Ο αλγόριθμος αυτός σχεδιάστηκε για την επίλυση δύο βασικών αδυναμιών των κλασικών τυχαίων δασών: τη μεροληψία υπέρ των blocks με μεγαλύτερο πλήθος μεταβλητών και την αδυναμία αξιολόγησης της πληροφοριακής αξίας μεταξύ blocks ίσου μεγέθους. Θεωρητικά, είναι προτιμότερη η επιλογή μεταβλητών από μικρότερα blocks, καθώς αυτό συχνά υποδηλώνει υψηλότερη πυκνότητα προγνωστικής πληροφορίας. Τα προβλήματα αυτά αντιμετωπίζονται ορίζοντας την πιθανότητα δειγματοληψίας μιας μεταβλητής που ανήκει στο $m-$οστό block μέσω της ποσότητας v_m υπό τον περιορισμό $\sum_{m=1}^{M} p_m v_m = 1$. Οι παράμετροι v_m βελτιστοποιούνται αυτόματα. Για τον καθορισμό του αριθμού των μεταβλητών που θα εξεταστούν σε

κάθε διαχωρισμό (mtry), ακολουθείται η εξής διαδικασία: κάθε μεταβλητή επιλέγεται τυχαία η μία μετά την άλλη· αν μια μεταβλητή ξαναεπιλεγεί, τότε η διαδικασία συνεχίζει μέχρι την εμφάνιση μη επιλεγμένης μεταβλητής στο σχεδιασμό. Η διαδικασία ολοκληρώνεται με την επιλογή mtry μεταβλητών. Η τιμή του $mtry = \sum_{m=1}^{M} \sqrt{p_m}$.

- **SplitWeights: Βάρη τιμών κριτηρίων διαχωρισμού για συγκεκριμένα block**

Σε αντίθεση με τη VarProb, η οποία επηρεάζει έμμεσα την ιεράρχηση των blocks μέσω της δειγματοληψίας, η SplitWeights παρεμβαίνει απευθείας στο κριτήριο διαχωρισμού. Ορίζονται ειδικά βάρη $w_1, \dots, w_M > 0$ για κάθε block m, με τον περιορισμό ταυτοποίησης $\max\{w_1, \dots, w_M\} = 1$. Με τον τρόπο αυτό, μεταβλητές που ανήκουν σε blocks με υψηλότερα βάρη αποκτούν προτεραιότητα διαχωρισμού. Κατά την υλοποίηση, επιλέγεται αρχικά ένας αριθμός μεταβλητών $mtry = \sum_{m=1}^{M} \sqrt{p_m}$ από όλες τις μεταβλητές όπως στο RSF. Σε δεύτερο στάδιο, υπολογίζονται οι τιμές του κριτηρίου διαχωρισμού σε όλα τα σημεία και σταθμίζονται με τα αντίστοιχα βάρη w_m του εκάστοτε block. Τέλος, επιλέγεται το σημείο διαχωρισμού που μεγιστοποιεί τη σταθμισμένη αυτή τιμή. Ένα μειονέκτημα της μεθόδου είναι η τάση επιλογής μεταβλητών από μεγάλα blocks με αραιή πληροφορία κάτι που απαιτεί προσεκτική ρύθμιση των βαρών w_m.

- **BlockVarSel: Ξεχωριστή δειγματοληψία μεταβλητών από κάθε block και βάρη τιμών διαχωρισμού ειδικά για κάθε block**

Ο αλγόριθμος BlockVarSel αναπτύχθηκε για να αντιμετωπίσει το μειονέκτημα της μεθόδου SplitWeights, σύμφωνα με το οποίο οι μεταβλητές που ανήκουν σε blocks μικρότερης διάστασης έχουν μικρότερη πιθανότητα να συμπεριληφθούν στο τυχαίο δείγμα υποψήφιων μεταβλητών για έναν διαχωρισμό. Στο BlockVarSel, αντί να πραγματοποιείται ενιαία δειγματοληψία από το σύνολο των μεταβλητών, πραγματοποιείται ξεχωριστή δειγματοληψία από κάθε block σε κάθε κόμβο, με τον αριθμό των επιλεγόμενων μεταβλητών να καθορίζεται από το μέγεθος του αντίστοιχου block. Στη συνέχεια, η επιλογή του σημείου διαχωρισμού πραγματοποιείται όπως και στη SplitWeights, χρησιμοποιώντας σταθμισμένες τιμές του κριτηρίου διαχωρισμού με βάρη ειδικά για κάθε block. Με τον τρόπο αυτό, διασφαλίζεται ότι μεταβλητές από κάθε block έχουν τη δυνατότητα να συμμετέχουν στη διαδικασία επιλογής διαχωρισμού, ενώ ταυτόχρονα από blocks μεγαλύτερης διάστασης επιλέγεται μεγαλύτερος αριθμός μεταβλητών.

- **RandomBlock: Τυχαία επιλογή block**

Βασιζόμενος στην υψηλή προγνωστική ικανότητα των RSF, ο αλγόριθμος RandomBlock εισάγει ένα επιπλέον επίπεδο τυχαιότητας. Σε κάθε κόμβο, επιλέγεται τυχαία ένα block και έπειτα λαμβάνεται από το συγκεκριμένο block ένα υποσύνολο μεταβλητών. Αυτό το υποσύνολο λαμβάνεται υπόψη για διαχωρισμό όπως στα RSF. Για να ενσωματώσουν την προγνωστική πληροφορία κάθε block ορίζεται πιθανότητα επιλογής ανά block b_m όπου $\sum_{m=1}^{M} b_m = 1$. Μετά την επιλογή block, $\sqrt{p_m}$

μεταβλητές δειγματίζονται από το συγκεκριμένο block. Με το συγκεκριμένο αλγόριθμο, αποφεύγονται προβλήματα διαστατικότητας μεταξύ των block και προβλήματα συσχετίσεων μεταξύ μεταβλητών από διαφορετικά block. Τέλος, οι τιμές b_m μπορούν να παίξουν καθοριστικό ρόλο στη σημασία διαφορετικών blocks για πρόβλεψη. Η επιλογή των b_m είναι σημαντική καθώς μεγάλες τιμές υποδεικνύουν συσχέτιση προγνωστικής πληροφορίας με άλλα blocks.

- **BlockForest: Δειγματοληψία block με ξεχωριστή δειγματοληψία μεταβλητών από κάθε block και βάρη ανά block των τιμών κριτηρίων που διαχωρίζονται**

Ο συγκεκριμένος αλγόριθμος αποτελεί συνδυασμό των RandomBlock και BlockVarSel.

1) Επιλέγεται ένα υποσύνολο όλων των M block, επιλέγοντας κάθε block με πιθανότητα 0.5, δηλαδή όλα τα blocks επιλέγονται με πιθανότητα 0.5^M και κανένα block με την ίδια πιθανότητα.
2) Αν δεν επιλεχθεί κανένα block επαναλαμβάνεται το πρώτο βήμα μέχρι την επιλογή τουλάχιστον ενός block.
3) Εφαρμογή του αλγορίθμου BlockVarSel στα επιλεγμένα blocks των βημάτων 1 ή 2.

- **Βελτιστοποίηση των παραμέτρων συντονισμού**

Για κάθε παράμετρο συντονισμού που προτείνεται στους παραπάνω αλγορίθμους εκτελείται για καθεμία ένας αλγόριθμος βελτιστοποίησης αυτών για τα καλύτερα δυνατά αποτελέσματα. Ο αλγόριθμος βελτιστοποίησης έχει ως εξής:

1) Για $it = 1, \dots N_{sets}$:
   a) Δημιουργείται ένα τυχαίο σύνολο $\boldsymbol{S}_{it}$ από M τιμές παραμέτρων συντονισμού
   b) Κατασκευάζεται ένα δάσος με $num.trees_{pre}$ δέντρα χρησιμοποιώντας το σύνολο τιμών παραμέτρων συντονισμού $\boldsymbol{S}_{it}$
2) Χρησιμοποιείται αυτό το σύνολο παραμέτρων συντονισμού $\boldsymbol{S}_1, \dots, \boldsymbol{S}_{N_{sets}}$ για το οποίο το αντίστοιχο δάσος παρείχε το μικρότερο σφάλμα πρόβλεψης εκτός σάκου (bag)

Για περισσότερες πληροφορίες αναφορικά με τον τρόπο επιλογής των τυχαίων συνόλων $\boldsymbol{S}_{it}$ δίνονται στην εργασία των Hornung & Wright (2019). Το μέτρο σφάλματος πρόβλεψης που συνίσταται είναι ο δείκτης C του Harrell. Το βασικό πλεονέκτημα του παραπάνω αλγορίθμου βελτιστοποίησης δεν είναι η αποτελεσματικότητα του, αλλά η συνέπεια σε σχέση με το πραγματικό σύνολο τιμών των παραμέτρων συντονισμού που ελαχιστοποιεί το σφάλμα πρόβλεψης εκτός σάκου (bag). Η συγκεκριμένη διαδικασία μπορεί να οδηγήσει και σε τοπικά βέλτιστα.

Οι Hornung & Wright (2019) έδειξαν, μέσω προσομοιώσεων, ότι ο αλγόριθμος BlockForest έχει σημαντικά καλύτερη απόδοση από τα RSF. Οι δύο προτεινόμενες παραλλαγές, που είχαν ενσωματώσει τυχαιότητα, είχαν την καλύτερη απόδοση, ενώ

στην περίπτωση συνόλου δεδομένων με ένα τύπο ωμικών δεδομένων, οι αλγόριθμοι καταλήγουν σε ακόμη καλύτερα αποτελέσματα από ότι στα πολυωμικά δεδομένα.

3.4 Αλγόριθμος Ενίσχυσης

Στο γενικό πλαίσιο των μεθόδων συνόλων μάθησης (ensemble learning), πέρα από τα Τυχαία Δάση Επιβίωσης και τις διαδικασίες bagging, εντάσσονται και οι αλγόριθμοι ενίσχυσης (boosting algorithms). Η αρχική ιδέα των αλγορίθμων ενίσχυσης προτάθηκε από τους Freund & Schapire (1997), οι οποίοι ανέπτυξαν τον αλγόριθμο AdaBoost. Πρόκειται για μία μέθοδο που εφαρμόζεται σε προβλήματα ταξινόμησης και παλινδρόμησης και βασίζεται στη διαδοχική προσαρμογή «αδύναμων μαθητών», στους οποίους αποδίδονται κατάλληλα βάρη. Με αυτόν τον τρόπο, μεγαλύτερη έμφαση δίνεται στις παρατηρήσεις που παρουσιάζουν μεγαλύτερη δυσκολία πρόβλεψης στα προηγούμενα στάδια της διαδικασίας. Η διαδικασία αυτή επαναλαμβάνεται διαδοχικά, με αποτέλεσμα τη δημιουργία ενός τελικού μοντέλου πρόβλεψης, το οποίο προκύπτει από το συνδυασμό των επιμέρους αδύναμων μαθητών που δημιουργήθηκαν κατά την εκτέλεση του αλγορίθμου. Μέσω αυτής της προσέγγισης επιτυγχάνεται η βελτίωση της προγνωστικής απόδοσης του μοντέλου και η αποτελεσματικότερη αντιμετώπιση σύνθετων προβλημάτων πρόβλεψης. Με βάση την ίδια φιλοσοφία, ο Ridgeway (1999) πρότεινε την ενσωμάτωση της μερικής συνάρτησης πιθανοφάνειας του μοντέλου Cox στον αλγόριθμο AdaBoost, με στόχο τη βελτίωση της προγνωστικής του ικανότητας. Η συγκεκριμένη προσέγγιση προσέλκυσε το ενδιαφέρον πολλών ερευνητών στο χώρο της Στατιστικής, οδηγώντας στην ανάπτυξη νέων τροποποιήσεων και επεκτάσεων της αρχικής μεθόδου. Μία σημαντική εφαρμογή των αλγορίθμων ενίσχυσης στην Ανάλυση Επιβίωσης παρουσιάστηκε από τους Binder & Schumacher (2008). Η εργασία τους αποτέλεσε τη βάση για την ανάπτυξη του πακέτου «CoxBoost» στη γλώσσα προγραμματισμού R και εισήγαγε έναν αλγόριθμο ενίσχυσης κατάλληλο για λογοκριμένα δεδομένα επιβίωσης. Η προτεινόμενη μέθοδος βασίζεται σε έναν μηχανισμό μετατόπισης, ο οποίος χρησιμοποιείται για την ποινικοποίηση των επεξηγηματικών μεταβλητών και τη δημιουργία απλούστερων μοντέλων με ικανοποιητική προγνωστική ικανότητα.

Μέχρι εκείνη την περίοδο, οι περισσότερες τεχνικές ενίσχυσης που έχουν προταθεί βασίζονται σε κάποιο υφιστάμενο μοντέλο επιβίωσης, όπως το ημιπαραμετρικό μοντέλο Cox. Οι Chen et al. (2013) πρότειναν μία μη παραμετρική μέθοδο ενίσχυσης κλίσης (Gradient Boosting Method – GBM), η οποία βασίζεται σε ένα σύνολο δέντρων παλινδρόμησης και στοχεύει στη βελτιστοποίηση του δείκτη συμφωνίας (Concordance Index – C-Index). Ο συγκεκριμένος δείκτης χρησιμοποιείται ευρέως στην Ανάλυση Επιβίωσης για την αξιολόγηση της προγνωστικής απόδοσης ενός μοντέλου. Με βάση την εφαρμογή της μεθόδου σε πραγματικά δεδομένα, οι Chen et al. (2013) διαπίστωσαν ότι η συγκεκριμένη τεχνική (Gradient Boosting Method Concordance Index – GBMCI) παρουσιάζει καλύτερη απόδοση σε σύγκριση με το κλασικό μοντέλο Cox, το μοντέλο Cox με εφαρμογή ενίσχυσης καθώς και τα Τυχαία Δάση Επιβίωσης RSF. Ωστόσο, η μέθοδος παρουσιάζει ορισμένους περιορισμούς, οι οποίοι σχετίζονται με τη διαδικασία

αρχικοποίησης και τον αυξημένο χρόνο υλοποίησης. Ο τελευταίος οφείλεται στους απαιτητικούς ζευγαρωτούς υπολογισμούς που απαιτούνται κατά τη διαδικασία εκπαίδευσης του αλγορίθμου. Σε παρόμοιο πλαίσιο κινήθηκαν και οι Mayr & Schmid (2013), οι οποίοι πρότειναν ένα ενιαίο πλαίσιο για την κατασκευή και αξιολόγηση συνδυασμών βιοδεικτών για την πρόβλεψη εκβάσεων χρόνου έως το συμβάν. Η προτεινόμενη μεθοδολογία βασίζεται στο δείκτη συμφωνίας C – index, ο οποίος αποτελεί μη παραμετρικό μέτρο της διακριτικής ικανότητας ενός κανόνα πρόβλεψης. Συγκεκριμένα, ανέπτυξαν έναν αλγόριθμο ενίσχυσης κλίσης, μέσω του οποίου προκύπτουν γραμμικοί συνδυασμοί βιοδεικτών που βελτιστοποιούν μια ομαλοποιημένη εκδοχή του C – index. Η απόδοση της μεθόδου αξιολογήθηκε και οι Mayr & Schmid (2013) έδειξαν ότι η προτεινόμενη προσέγγιση μπορεί να επιτύχει υψηλότερη διακριτική ικανότητα σε σύγκριση με παραδοσιακές μεθόδους κατασκευής γονιδιακών υπογραφών.

Σε περιπτώσεις όπου ο αριθμός των επεξηγηματικών μεταβλητών είναι σημαντικά μεγαλύτερος από τον αριθμό των διαθέσιμων παρατηρήσεων, η αντιμετώπιση όλων των μεταβλητών με τον ίδιο βαθμό αραιότητας δεν είναι πάντοτε αποτελεσματική. Για τον λόγο αυτό, οι Sariyar et al. (2014) πρότειναν μια προσαρμοσμένη μέθοδο ενίσχυσης (AdaptiveBoost), στην οποία κάθε μεταβλητή λαμβάνει ένα συγκεκριμένο βάρος. Το βάρος αυτό ενημερώνεται κατάλληλα όταν η αντίστοιχη μεταβλητή επιλέγεται σε ένα εσωτερικό βήμα του αλγορίθμου ενίσχυσης. Σημαντικό πλεονέκτημα της συγκεκριμένης μεθόδου αποτελεί η αντιμετώπιση της υποεκτίμησης των συντελεστών, καθώς και η δημιουργία ενός αυτοματοποιημένου μηχανισμού επιβολής αραιότητας στα δεδομένα. Οι Sariyar et al. (2014) ανέπτυξαν μια ολοκληρωμένη προσέγγιση βασισμένη στην πιθανοφάνεια ανά συνιστώσα, μέσω της οποίας αντιμετώπισαν προβλήματα που μέχρι τότε εξετάζονταν μεμονωμένα και πρότειναν μία τεχνική μείωσης της αρχική διάστασης των δεδομένων. Παρά τα πλεονεκτήματα της μεθόδου, η εφαρμογή της σε μη αραιά περιβάλλοντα δεδομένων παρουσιάζει ορισμένους περιορισμούς. Συγκεκριμένα, σε τέτοιες περιπτώσεις παρατηρείται αυξημένη ευαισθησία του αλγορίθμου, μεγαλύτερη μεταβλητότητα στις εκτιμήσεις των συντελεστών και, κατά συνέπεια, αύξηση των ψευδώς θετικών ανακαλύψεων (false positives discoveries). Αντίθετα, σε αραιά περιβάλλοντα δεδομένων εμφανίζεται μειωμένη μεροληψία στις εκτιμήσεις των συντελεστών, βελτιωμένη αναγνώριση των σημαντικών μεταβλητών, ικανοποιητική προγνωστική απόδοση και μεγαλύτερη σταθερότητα στη διαδικασία επιλογής τους. Επιπλέον, ιδιαίτερη σημασία έχει η κατάλληλη επιλογή των παραμέτρων συντονισμού, καθώς σημαντικές αποκλίσεις από τις βέλτιστες τιμές τους μπορούν να οδηγήσουν σε διαφορετικά αποτελέσματα και να επηρεάσουν σημαντικά την απόδοση του τελικού μοντέλου.

Η ανάγκη για αποτελεσματική μείωση της υψηλής διάστασης σε δεδομένα επιβίωσης αποτέλεσε αντικείμενο μελέτης πολλών ερευνητών και οδήγησε στην ανάπτυξη νέων υπολογιστικά αποδοτικών αλγορίθμων. Μία τέτοια προσέγγιση παρουσιάστηκε από τους He et al. (2016), οι οποίοι πρότειναν μια νέα τεχνική ενίσχυσης (gradient boosting with stability selection) για την ανάλυση δεδομένων

υψηλής διάστασης. Η προτεινόμενη μέθοδος αξιολογήθηκε και συγκρίθηκε με ευρέως χρησιμοποιούμενες μονομεταβλητές προσεγγίσεις και με τη μέθοδο Lasso για την επιλογή μεταβλητών. Τα αποτελέσματα έδειξαν ότι η προτεινόμενη προσέγγιση υπερέχει ως προς το σημαντικό περιορισμό των ψευδώς θετικών ευρημάτων (false positives), διατηρώντας παράλληλα χαμηλό αριθμό ψευδώς αρνητικών ευρημάτων (false negatives). Επιπλέον, οι He et al. (2016) επισημαίνουν ότι η τροποποιημένη διαδικασία ενίσχυσης κλίσης μπορεί να εφαρμοστεί σε πιο ευέλικτους χώρους παραμέτρων, ακόμη και σε άπειρης διάστασης συναρτησιακούς χώρους. Ωστόσο, η εφαρμογή σε τέτοιους χώρους απαιτεί τον υπολογισμό της παραγώγου Gâteaux για τον προσδιορισμό της βέλτιστης κατεύθυνσης καθόδου.

Την επιλογή σταθερότητας και τη βελτιστοποίηση του δείκτη συμφωνίας C – Index επιχείρησαν να βελτιώσουν οι Mayr et al. (2016) μέσω της προτεινόμενης μεθόδου τους. Συγκεκριμένα, συνδύασαν μια αυτοματοποιημένη διαδικασία επιλογής μεταβλητών με την προσαρμογή μοντέλων αραιής διάκρισης για υψηλής διάστασης δεδομένα επιβίωσης, χρησιμοποιώντας αλγόριθμο ενίσχυσης, υποδειγματοληψία και βελτιστοποίηση του δείκτη C – Index. Βασικός περιορισμός της συγκεκριμένης τεχνικής είναι η ανάγκη ύπαρξης ενός μικρού συνόλου ισχυρά προγνωστικών βιοδεικτών, ώστε να επιτυγχάνεται η μέγιστη δυνατή απόδοσή της και να μπορεί να παρουσιάσει καλύτερα αποτελέσματα σε σύγκριση με ποινικοποιημένα μοντέλα Cox, όπως το Cox – Lasso. Οι συγγραφείς κατέληξαν ότι δεν είναι εφικτή η ταυτόχρονη επίτευξη ενός μοντέλου με υψηλή ερμηνευσιμότητα (αραιό μοντέλο) και μεγάλης διακριτικής ικανότητας.

3.4.1 XGBLC Αλγόριθμος

Ο συνδυασμός του ακραίου αλγορίθμου ενίσχυσης κλίσης και του μοντέλου Lasso – Cox (eXtreme Gradient Boosting and Lasso Cox algorithm – XGBLC) αποτελεί επέκταση της μεθόδου XGBoost των Chen & Guestrin (2016). Η συγκεκριμένη μέθοδος αναπτύχθηκε από τους Ma et al. (2022) και έχει παρουσιάσει σημαντικό ενδιαφέρον στο χώρο της Ανάλυσης Επιβίωσης, καθώς προτείνει ένα βελτιωμένο μοντέλο πρόβλεψης επιβίωσης με δυνατότητα μείωσης της αρχικής διάστασης των δεδομένων. Η ιδιότητα αυτή είναι ιδιαίτερα σημαντική σε εφαρμογές ιατρικών δεδομένων, καθώς μπορεί να συμβάλει στην έγκαιρη και αξιόπιστη λήψη κλινικών αποφάσεων. Η ανάπτυξη της συγκεκριμένης τεχνικής βασίστηκε στην αξιοποίηση γονιδιακών δεδομένων, τα οποία χαρακτηρίζονται συνήθως από πολύ μεγάλο αριθμό επεξηγηματικών μεταβλητών, οι οποίες μπορεί να ανέρχονται σε δεκάδες χιλιάδες γονίδια. Η ύπαρξη δεδομένων υψηλής διάστασης δημιουργεί σημαντικές προκλήσεις τόσο στη διαδικασία εκτίμησης των παραμέτρων όσο και στην επιλογή των μεταβλητών που παρουσιάζουν πραγματική προγνωστική αξία. Για την αντιμετώπιση των παραπάνω προβλημάτων, οι Ma et al. (2022) συνδύασαν την προγνωστική ικανότητα του αλγορίθμου XGBoost με τη διαδικασία επιλογής μεταβλητών που προσφέρει η ποινή Lasso στο μοντέλο Cox. Με τον τρόπο αυτό επιτυγχάνεται η ταυτόχρονη μείωση του θορύβου στα δεδομένα και η αναγνώριση των σημαντικότερων επεξηγηματικών μεταβλητών.

Έστω ότι υπάρχει ένα σύνολο εκπαίδευσης: $D = \{(\boldsymbol{X}_i, y_i, \delta_i)\}$ με n παρατηρήσεις και m γονιδιακά χαρακτηριστικά (επεξηγηματικές μεταβλητές). Το $\boldsymbol{X}_i$ αντιστοιχεί στο παρατηρούμενο διάνυσμα χαρακτηριστικών του $i-$οστού ατόμου, το y_i στον παρατηρούμενο χρόνο επιβίωσης και το $\delta_i = 0$ δηλώνει ότι το $i-$οστό άτομο είναι λογοκριμένο από δεξιά. Ο αλγόριθμος XGBoost των Chen & Guestrin (2016) χρησιμοποιεί Κ πρόσθετες συναρτήσεις για την εκτίμηση του προβλεπόμενου χρόνου επιβίωσης $\hat{y_i}$, σύμφωνα με τη σχέση:

$$\hat{y_i} = \sum_{j=1}^{K} f_j(X_i), f_j \in F \quad \text{όπου}$$

- $f(X)$: η πρόβλεψη του αλγορίθμου CART που χρησιμοποιούνται στον αλγόριθμο XGBoost.
- $F = \{f(X) = \omega_{q(x)}\}(q: \mathbb{R}^m \to T, \omega \in R^T)$ αποτελεί το χώρο των CART δέντρων που επιτρέπεται να χρησιμοποιήσει ο αλγόριθμος XGBoost.
- q: αφορά τη δομή του εκάστοτε δέντρου, δηλαδή την τοπολογία του δέντρου.
- T: είναι ο αριθμός των φύλλων (leaves) κάθε δέντρου.

Κάθε f_j αντιπροσωπεύει ένα ανεξάρτητο δέντρο με δομή q ενώ οι κόμβοι του προσαρμόζονται μέσω των αντίστοιχων παραμέτρων βάρους $\boldsymbol{\omega}$. Η διαδικασία εκπαίδευσης των δέντρων πραγματοποιείται μέσω της ελαχιστοποίησης της κανονικοποιημένης συνάρτησης απώλειας:

$$L^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \Omega(f_t)$$

όπου l είναι μια διαφορίσιμη κυρτή συνάρτηση απώλειας, η οποία μετρά τη διαφορά μεταξύ της πρόβλεψης $\hat{y}_i$ και της παρατηρούμενης τιμής y_i, t δηλώνει την επανάληψη και το $\Omega(f_t)$ εκφράζει τον περιορισμό που επιβάλλεται στο f_t. Η παραπάνω συνάρτηση μπορεί να προσεγγιστεί χρησιμοποιώντας τη ποινικοποιημένη μερική συνάρτηση πιθανοφάνειας του μοντέλου Cox με την προσέγγιση Breslow L_B και ποινή Lasso $P(B)$, δηλαδή $L_p = L_B - P(B)$ (βλ. τύπο (1.17) σελ. 21 και ποινή Lasso σελ. 49), λαμβάνοντας τη μορφή:

$$L^{(t)} \approx \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + g_i f_t(X_i) + \frac{1}{2} b_i f_t^2(X_i)\right) + \Omega(f_t) \qquad \text{όπου}$$

- g_i: εκφράζει τη στατιστική διαβάθμισης πρώτης τάξης της μερικής συνάρτησης πιθανοφάνειας L_p η οποία δίνεται από τον κάτωθι τύπο:

$$g_i = \frac{\partial L_p}{\partial \hat{y}_i} = \begin{cases} e^{\hat{y}_i}\left(\sum_{t \le T_i} \frac{1}{\sum_{j \in q(t)} S(t)}\right) - 1 - \lambda \sum_k \frac{1}{X_k}, & \text{εάν } \delta_i = 1 \\ e^{\hat{y}_i}\left(\sum_{t \le T_i} \frac{1}{\sum_{j \in q(t)} S(t)}\right) - \lambda \sum_k \frac{1}{X_k}, & \text{εάν } \delta_i = 0 \end{cases}$$

- b_i: εκφράζει τη στατιστική διαβάθμισης δεύτερης τάξης της μερικής συνάρτησης πιθανοφάνειας L_p η οποία δίνεται από τον τύπο:

$$b_i = \frac{\partial L_p^2}{\partial \hat{y}_i^2} = e^{\hat{y}_i}\left(\sum_{t \le T_i} \frac{1}{\sum_{j \in q(t)} S(t)}\right) - \left(e^{\hat{y}_i}\right)^2 \times \sum_{t \le T_i}\left(\frac{1}{S(t)}\right)^2$$

Έχοντας ορίσει όλες τις παραπάνω ποσότητες, η διαδικασία λειτουργίας του αλγορίθμου παρουσιάζεται στην Εικόνα 3.1.

Algorithm 1 Calculating the Gradient Statistics of Loss Function

Input: $\hat{y}$, predictions of individuals on log hazard ratio scale; δ, t, the survival data of individuals; X, the genomic features of individuals; λ, the penalty parameter of XGBLC algorithm
Output: *grad, hess*
 # data preparation and initialization
$T = \text{order}(t)$ *# sort by time*
$r_k = 0$
$s_k = 0$
$ST = \sum_{i \in R(t)} e^{\hat{y}_i}$
$\beta = (X^T X)^{-1} X^T y$
 # calculate each individual separately
for i in T
 # the Breslow approximation
    if $\delta_i > 0$
    $r_k = r_k + 1/S$
    $s_k = s_k + 1/(S)^2$
    $x = 1/X[i,]$ *# the i-th individual's gene expression data*
        if $\delta_i > 0$
        $\text{grad}_i = e^{\hat{y}_i} \times r_k - 1 - \lambda \times x$
        else if $\delta_i = 0$
            $grad_i = e^{\hat{y}_i} \times r_k - \lambda \times x$
        $hess_i = e^{\hat{y}_i} \times s_k - (e^{\hat{y}_i})^2 \times s_k$
return *grad, hess*

Εικόνα 3.1. Αλγόριθμος XGBoost : Εφαρμογή στην εργασία των Ma et al. (2022)

Η ενσωμάτωση του μοντέλου Cox με ποινή Lasso στον αλγόριθμο ενίσχυσης του XGBoost οδήγησε στη μείωση του θορύβου που υπήρχε στα δεδομένα, καθώς και στην αποτελεσματικότερη ανίχνευση των στατιστικά σημαντικών μεταβλητών. Με τον τρόπο αυτό ενισχύθηκε η προγνωστική ακρίβεια του τελικού μοντέλου. Για την αξιολόγηση της απόδοσης του προτεινόμενου αλγορίθμου πραγματοποιήθηκαν προσομοιώσεις, καθώς και εφαρμογή σε πραγματικό σύνολο δεδομένων υψηλής διάστασης, με στόχο τη σύγκρισή του με άλλες βασικές τεχνικές. Τα αποτελέσματα των προσομοιώσεων έδειξαν ότι η προτεινόμενη μέθοδος παρουσιάζει υψηλότερα επίπεδα ακρίβειας σε σχέση με τις υπόλοιπες τεχνικές που εξετάστηκαν. Ειδικότερα, από τα είκοσι σύνολα δεδομένων που χρησιμοποιήθηκαν, η μέθοδος παρουσίασε ικανοποιητική απόδοση στα δώδεκα από αυτά. Η μειωμένη απόδοση που παρατηρήθηκε στα υπόλοιπα οκτώ σύνολα δεδομένων αποδόθηκε κυρίως στην ύπαρξη ελλειπουσών τιμών, καθώς και στη διαφορετική διαδικασία επιλογής των κλινικών μεταβλητών. Παρά τη συνολικά βελτιωμένη συμπεριφορά της μεθόδου, ένας σημαντικός περιορισμός αφορά το χρόνο υλοποίησης, οποίος είναι αυξημένος σε σύγκριση με άλλους αλγορίθμους. Ο αυξημένος υπολογιστικός χρόνος οφείλεται στο μεγαλύτερο πλήθος παραμέτρων συντονισμού, καθώς και στην εισαγωγή της ποινής Lasso στη συνάρτηση απώλειας.

3.4.2 Αποδοτική ενίσχυση κλίσης

Στο χώρο της Στατιστικής και ειδικότερα στον τομέα της Μηχανικής Μάθησης έχουν αναπτυχθεί αρκετές τεχνικές που βασίζονται στα Δέντρα Απόφασης, στα Τυχαία Δάση και στους αλγορίθμους ενίσχυσης. Οι συγκεκριμένες μέθοδοι έχουν συμβάλει σημαντικά στην αντιμετώπιση προβλημάτων υψηλής διάστασης, τα οποία εμφανίζονται συχνά στην ανάλυση δεδομένων επιβίωσης. Στο πλαίσιο των αλγορίθμων ενίσχυσης έχει αναπτυχθεί ένα ευρύ σύνολο μεθόδων, με στόχο τη βελτίωση της προγνωστικής απόδοσης και την αποτελεσματικότερη διαχείριση μεγάλου αριθμού επεξηγηματικών μεταβλητών. Χαρακτηριστικά παραδείγματα αποτελούν η μέθοδος XGBoost των Chen & Guestrin (2016), καθώς και η Αραιή Ενίσχυση Κλίσης (Light Gradient Boosting – LGB) των Ke et al. (2017). Ωστόσο, οι συγκεκριμένες προσεγγίσεις δεν είχαν αρχικά σχεδιαστεί για εφαρμογή σε λογοκριμένα δεδομένα επιβίωσης.

Το παραπάνω πρόβλημα αντιμετωπίστηκε από τους Li et al. (2022), οι οποίοι πρότειναν μία προσαρμοσμένη μέθοδο ενίσχυσης κλίσης κατάλληλη για δεδομένα επιβίωσης. Η προτεινόμενη προσέγγιση συνδυάζει τη λογική των σύγχρονων αλγορίθμων ενίσχυσης με τη δομή των μοντέλων επιβίωσης, επιτρέποντας την αποτελεσματική εφαρμογή τους σε δεδομένα με παρουσία λογοκριμένων παρατηρήσεων.

Έστω ότι η συνάρτηση κινδύνου ορίζεται από την ακόλουθη σχέση:

$$\lambda(t|\boldsymbol{x}) = \lambda_0(t)\exp\{H(\boldsymbol{x}_i)\}$$

όπου η συνάρτηση $H(\cdot)$ αποτελεί μια συνάρτηση βαθμολογίας κινδύνου, η οποία συνδέει τις συμμεταβλητές με τους αντίστοιχους συντελεστές παλινδρόμησης. Στην περίπτωση του μοντέλου Cox, η συγκεκριμένη συνάρτηση εκφράζεται ως ο γραμμικός συνδυασμός $\boldsymbol{x}_i^T\boldsymbol{\beta}$, ενώ η εκτίμηση των συντελεστών πραγματοποιείται μέσω της ελαχιστοποίησης του ποινικοποιημένου αρνητικού λογαρίθμου της μερικής συνάρτησης πιθανοφάνειας:

$$\mathcal{L} = -\log PL = -\sum_i \delta_i \left\{ H(\boldsymbol{x}_i) - \log\left(\sum_{k\in R} \exp\big(H(\boldsymbol{x}_k)\big) \right) \right\}$$

όπου η ποινή καθορίζεται από τον εκάστοτε αναλυτή και μπορεί να αντιστοιχεί μεταξύ άλλων, στις μεθόδους Lasso, Ridge, Group Lasso, κ.α..

Η κλίση της συνάρτησης απώλειας Cox δίνεται από:

$$g_i = \delta_i - \sum_{j=1}^{n} \delta_j I\big(t_i \geq t_j\big) \frac{\exp\big(H(\boldsymbol{x}_i)\big)}{\sum_k^n I\big(t_k \geq t_j\big)\exp\big(H(\boldsymbol{x}_k)\big)}$$

Ο κλασικός αλγόριθμος ενίσχυσης κλίσης βελτιστοποιεί επαναληπτικά την παραπάνω συνάρτηση, επιλέγοντας σε κάθε στάδιο έναν «ασθενή μαθητή» (weak learner). Η βέλτιστη βασική συνάρτηση στο m – οστό βήμα υπολογίζεται μέσω της σχέσης:

$$\underset{\eta \in H}{arg\min} \sum_{i=1}^{n} \left(\eta(\boldsymbol{x}_i) - g_i^{(m)}\right)^2$$

Οι σύγχρονες προσεγγίσεις ενίσχυσης κλίσης αξιοποιούν επιπλέον τη δεύτερη παράγωγο της συνάρτησης απώλειας, η οποία μπορεί να θεωρηθεί ως μία βελτιωμένη μορφή της διαδικασίας boosting. Με τη χρήση της δεύτερης παραγώγου επιτυγχάνεται ακριβέστερη προσέγγιση της συνάρτησης απώλειας και βελτιώνεται η διαδικασία εκπαίδευσης του μοντέλου. Η εμπειρική συνάρτηση ενίσχυσης μπορεί να γραφεί ως:

$$g_i = \delta_i - \sum_{j=1}^{n} \delta_j \tilde{\pi}_{ij}$$

όπου

$$\tilde{\pi}_{ij} = I\left(t_i \geq t_j\right) \frac{\exp\left(H(\boldsymbol{x}_i)\right)}{\sum_k^n I\left(t_k \geq t_j\right) \exp\left(H(\boldsymbol{x}_k)\right)}$$

και η δεύτερη παράγωγος διαμορφώνεται ως:

$$s_i = -\sum_{j=1}^{n} \delta_j \, \tilde{\pi}_{ij}\left(1 - \tilde{\pi}_{ij}\right)$$

Με βάση τους παραπάνω ορισμούς, η συνάρτηση απώλειας $\boldsymbol{\mathcal{L}}$, στο m – οστό βήμα της διαδικασίας εκπαίδευσης, μπορεί να προσεγγιστεί από τη σχέση:

$$\mathcal{L}^{(\mathrm{m})} \approx \sum_{i=1}^{n} \left\{ g_i^{(m)} \eta^{(m)}(\boldsymbol{x}_i) + \frac{1}{2} s_i^{(m)} \left[\eta^{(m)}(\boldsymbol{x}_i)\right]^2 \right\} + C^{(m)}$$

όπου ο όρος C περιλαμβάνει μόνο τον όρο απώλειας του προηγούμενου βήματος $H^{(m-1)}$.

Η παραπάνω μορφή της συνάρτησης απώλειας είναι ισοδύναμη με ένα πρόβλημα παλινδρόμησης σταθμισμένων ελαχίστων τετραγώνων. Επομένως, η βέλτιστη βασική συνάρτηση προκύπτει από:

$$\eta^{(m)} = \underset{\eta \in H}{\operatorname{argmin}} \sum_{i=1}^{n} \left(\frac{1}{2} s_i^{(m)} \left[-\frac{g_i^{(m)}}{s_i^{(m)}} - \eta^{(m)}(\boldsymbol{x}_i) \right]^2 \right)$$

Η ενημερωμένη συνάρτηση μετά το βήμα (m) δίνεται από:

$$H^{(m)} = H^{(m-1)} + \nu \eta^{(m)}$$

όπου ν: αποτελεί την παράμετρο μάθησης. Στο πλαίσιο της ενίσχυσης δέντρων, ως βασικοί μαθητές χρησιμοποιούνται τα Δέντρα Απόφασης. Κάθε κόμβος του δέντρου αντιστοιχεί σε ένα συγκεκριμένο βάρος, ενώ η συνάρτηση βάσης του δέντρου ορίζεται ως:

$$\eta^{(m)}(\boldsymbol{x}) = \sum_{l=1}^{L} w_l \, I(\boldsymbol{x} \in T_l)$$

όπου w_l αποτελεί το βάρος του $l-$ οστού φύλλου και η δίτιμη συνάρτηση $I(\boldsymbol{x} \in T_l)$ καθορίζει τη δομή του αντίστοιχου δέντρου. Με βάση τη συγκεκριμένη αναπαράσταση, η εμπειρική συνάρτηση απώλειας μπορεί να εκφραστεί ως:

$$\mathcal{L}^{(\mathrm{m})} = \sum_{l=1}^{L} \sum_{i \in I_l} \left\{ g_i^{(m)} w_l + \frac{1}{2} s_i^{(m)} w_l^2 \right\} = \sum_{l=1}^{L} \left\{ G_l^{(m)} w_l + \frac{1}{2} S_l^{(m)} w_l^2 \right\}$$

όπου G_l και S_l αποτελούν τα αθροίσματα των g_i και s_i, αντίστοιχα, για τον κάθε κόμβο του δέντρου. Ακολουθεί η διαδικασία βελτιστοποίησης των βαρών των κόμβων, από την οποία προκύπτει η εκτίμηση:

$$\widehat{w}_l = -G_l^{(m)} / S_l^{(m)}$$

ενώ η αντίστοιχη βέλτιστη συνάρτηση σκορ δίνεται από:

$$-\frac{1}{2} \sum_{l=1}^{L} \frac{\left[G_l^{(m)} \right]^2}{S_l^{(m)}}$$

Για τον περιορισμό της πολυπλοκότητας του μοντέλου και την αποφυγή φαινομένων υπερπροσαρμογής, εισάγεται μία ποινή στη συνάρτηση απώλειας. Με την προσθήκη της συγκεκριμένης ποινής, η βέλτιστη συνάρτηση σκορ διαμορφώνεται ως:

$$-\frac{1}{2} \sum_{l=1}^{L} \frac{\left[G_l^{(m)} \right]^2}{S_l^{(m)} + \lambda} + \gamma L$$

όπου λ: η παράμετρος κανονικοποίησης της ποινής l_2, ενώ γ: η παράμετρος ποινής που ελέγχει το πλήθος των κόμβων του δέντρου. Η μείωση της συνάρτησης απώλειας μετά από μια διαδικασία διαχωρισμού (split), το κέρδος – σκορ (gain score), υπολογίζεται ως:

$$L_{split} = \frac{1}{2} \left[\frac{G_L^2}{S_L + \lambda} + \frac{G_R^2}{S_R + \lambda} - \frac{(G_L + G_R)^2}{S_L + S_R + \lambda} \right] - \gamma$$

Με βάση την παραπάνω διαδικασία, οι Li et al. (2022) κατάφεραν να απλοποιήσουν τη διαδικασία αναζήτησης των βέλτιστων διαχωρισμών, ελαχιστοποιώντας απευθείας την εμπειρική συνάρτηση μέσω της παραγώγου της συνάρτησης κέρδους. Ωστόσο, η εφαρμογή της συγκεκριμένης μεθόδου προϋποθέτει την ισχύ της υπόθεσης αναλογικών κινδύνων, η οποία αποτελεί βασική προϋπόθεση των μοντέλων Cox. Για την αντιμετώπιση αυτού του περιορισμού, οι Li et al. (2022) πρότειναν μία εναλλακτική προσέγγιση βασισμένη στη βελτιστοποίηση του δείκτη συμφωνίας C – Index. Η γενική λογική της μεθόδου παραμένει ίδια, με τη διαφορά ότι η αρχική συνάρτηση απώλειας αντικαθίσταται από την ακόλουθη:

$$SCI = \frac{1}{|P|} \sum_{(i,j) \in P} \frac{1}{1 + \exp\left(a(H(\boldsymbol{x}_i) - \left(H(\boldsymbol{x}_j) \right) \right)}$$

όπου η υπερπαράμετρος α χρησιμοποιείται για τη ρύθμιση της κλίσης της συνάρτησης.

Τα αποτελέσματα των προσομοιώσεων που πραγματοποιήθηκαν από τους Li et al. (2022), έδειξαν ότι η προτεινόμενη μέθοδος παρουσιάζει υψηλή προγνωστική ακρίβεια και αποτελεσματική επιλογή βιοδεικτών. Συγκεκριμένα, η μέθοδος LGB εμφάνισε καλύτερη συμπεριφορά σε δεδομένα ιδιαίτερα υψηλής διάστασης σε σύγκριση με τη μέθοδο XGBoost, παρουσιάζοντας μεγαλύτερη αποτελεσματικότητα στη διαχείριση μεγάλου αριθμού επεξηγηματικών μεταβλητών. Επιπλέον, μέσω της συνολικής μεθοδολογίας επιτεύχθηκε η οπτικοποίηση ενός απλού Δέντρου Απόφασης, το οποίο μπορεί να αξιοποιηθεί για την υποστήριξη σημαντικών ιατρικών αποφάσεων. Παράλληλα, οι συγγραφείς ανέπτυξαν το πακέτο «XSurv» στη γλώσσα προγραμματισμού R, το οποίο περιλαμβάνει τις απαραίτητες διαδικασίες και ρουτίνες για την εφαρμογή της μεθόδου και παρέχει τη δυνατότητα βαθμονόμησης του χρόνου επιβίωσης. Ιδιαίτερα σημαντική είναι και η διαδικασία προσδιορισμού των βέλτιστων υπερπαραμέτρων του μοντέλου. Οι Li et al. (2022) προτείνουν τη χρήση διασταυρούμενης επικύρωσης (CV) για την κατάλληλη επιλογή τους, καθώς τιμές σημαντικά διαφορετικές από τις βέλτιστες μπορούν να οδηγήσουν σε διαφοροποίηση της συμπεριφοράς του αλγορίθμου και, συνεπώς, σε διαφορετικά αποτελέσματα πρόβλεψης. Συνολικά, οι αλγόριθμοι ενίσχυσης αποτελούν μία ιδιαίτερα σημαντική κατηγορία μεθόδων στην Ανάλυση Επιβίωσης, καθώς επιτρέπουν την αποτελεσματική διαχείριση σύνθετων και υψηλής διάστασης δεδομένων. Παρά το γεγονός ότι παρουσιάζουν αυξημένες υπολογιστικές απαιτήσεις και απαιτούν προσεκτική ρύθμιση των παραμέτρων τους, η δυνατότητά τους να συνδυάζουν υψηλή προγνωστική ακρίβεια με διαδικασίες επιλογής σημαντικών μεταβλητών τους καθιστά ιδιαίτερα χρήσιμους σε σύγχρονες εφαρμογές, όπως η ανάλυση γονιδιακών και κλινικών δεδομένων.

3.5 Μηχανή υποστήριξης διανυσμάτων

Οι μηχανές υποστήριξης διανυσμάτων (Support Vector Machines – SVM) αποτελούν μία από τις σημαντικότερες τεχνικές της Μηχανικής Μάθησης, η ανάπτυξη των οποίων ξεκίνησε τη δεκαετία του 1960, ενώ η βασική θεωρητική τους θεμελίωση πραγματοποιήθηκε τη δεκαετία του 1990 με στόχο την αντιμετώπιση μη γραμμικών προβλημάτων ταξινόμησης. Αρχικά, η συγκεκριμένη μέθοδος επιβλεπόμενης μάθησης (supervised learning method) εφαρμόστηκε σε προβλήματα ταξινόμησης, ενώ στη συνέχεια επεκτάθηκε και σε μοντέλα παλινδρόμησης. Αποτελεί μία από τις πλέον διαδεδομένες τεχνικές στο χώρο της Στατιστικής, καθώς μπορεί να διαχειριστεί αποτελεσματικά μεγάλο αριθμό επεξηγηματικών μεταβλητών ακόμη και όταν το πλήθος των διαθέσιμων παρατηρήσεων είναι περιορισμένο, χωρίς ωστόσο η εφαρμογή της να περιορίζεται αποκλειστικά σε τέτοιου είδους προβλήματα. Η βασική ιδέα της μεθόδου στηρίζεται στη χαρτογράφηση των παρατηρήσεων σε έναν χώρο, όπου προσδιορίζεται ένα υπερεπίπεδο μεγίστου περιθωρίου (maximum margin hyperplane) που επιτυγχάνει το βέλτιστο διαχωρισμό των υπό μελέτη ομάδων. Η διαδικασία αυτή πραγματοποιείται μέσω της χρήσης συναρτήσεων πυρήνα (kernel

functions), οι οποίες επιτρέπουν την αποτελεσματική αντιμετώπιση μη γραμμικών σχέσεων μεταξύ των μεταβλητών. Μετά την εκπαίδευση του μοντέλου, κάθε νέα παρατήρηση μπορεί να ταξινομηθεί στην κατάλληλη ομάδα με βάση το υπερεπίπεδο που έχει προκύψει. Τα σημαντικότερα πλεονεκτήματα της μεθόδου περιλαμβάνουν την ύπαρξη καθολικά βέλτιστης λύσης μέσω κυρτού προγραμματισμού, την ισχυρή ικανότητα γενίκευσης, τη δυνατότητα αξιοποίησης όλων των επεξηγηματικών μεταβλητών ακόμη και σε προβλήματα υψηλής διάστασης, καθώς και την ευρεία υποστήριξή της από τις περισσότερες γλώσσες προγραμματισμού. (Montesinos Lopez et al. (2022))

Τα SVM αναπτύχθηκαν αρχικά για προβλήματα στα οποία η μεταβλητή απόκρισης είναι δίτιμη, με στόχο την ταξινόμηση των παρατηρήσεων στις δύο υπό μελέτη ομάδες (Support Vector Classification – SVC). Στη συνέχεια, η μεθοδολογία επεκτάθηκε και σε προβλήματα παλινδρόμησης μέσω των μοντέλων Support Vector Regression (SVR), στα οποία η μεταβλητή απόκρισης είναι συνεχής και ο αλγόριθμος εκπαιδεύεται για την ανάπτυξη μοντέλων πρόβλεψης. Το 2007 ξεκίνησε η επέκταση των συγκεκριμένων τεχνικών και σε προβλήματα που περιλαμβάνουν λογοκριμένες παρατηρήσεις. Η πρώτη προσέγγιση για λογοκριμένα δεδομένα, με την ονομασία Support Vector Censored Regression (SVCR), προτάθηκε από τους Shivaswamy et al. (2007) και βασίζεται στην ενσωμάτωση των λογοκριμένων στόχων στο SVM μέσω κατάλληλα ορισμένων διαστημάτων. Με βάση τα πειραματικά αποτελέσματα που παρουσίασαν, η προτεινόμενη μέθοδος εμφάνισε καλύτερη απόδοση τόσο ως προς την κατάταξη όσο και ως προς την πρόβλεψη του χρόνου επιβίωσης, σε σύγκριση με τις κλασικές προσεγγίσεις της Ανάλυσης Επιβίωσης, όπως τα SVR, CC – SVM (Constraint Classification – Support Vector Machine), GP – PL (Gaussian Process Preference Learning) και τα κλασικά παραμετρικά μοντέλα.

Τα επόμενα έτη, αρκετοί ερευνητές ασχολήθηκαν με την τροποποίηση των SVM και SVR, με στόχο την αποτελεσματική επεξεργασία λογοκριμένων δεδομένων. Στο πλαίσιο αυτό, οι Evers & Messow (2008), βασιζόμενοι στην αρχική θεωρία των SVM και στη χρήση συναρτήσεων πυρήνα, πρότειναν την επέκταση των μεθόδων αραιού πυρήνα σε λογοκριμένα δεδομένα μέσω του ημιπαραμετρικού μοντέλου αναλογικών κινδύνων του Cox. Συγκεκριμένα, ανέπτυξαν δύο διαφορετικές τεχνικές. Η πρώτη, γνωστή ως Survival Support Vector Machine (Survival SVM), βασίζεται σε μία γεωμετρική προσέγγιση παρόμοια με εκείνη των SVC, όπου μεγιστοποιείται το περιθώριο μεταξύ της λογοκριμένης παρατήρησης και των παρατηρήσεων που βρίσκονται εκείνη τη χρονική στιγμή σε κίνδυνο. Η δεύτερη, με την ονομασία Survival Import Vector Machine (Survival IVM), δημιουργεί ένα αραιό μοντέλο προσθέτοντας διαδοχικά παρατηρήσεις, ακολουθώντας τη φιλοσοφία της Μηχανικής Εισαγωγής Διανυσμάτων των Zhu & Hastie (2005). Ωστόσο, οι συγκεκριμένες τεχνικές επιτυγχάνουν αραιότητα ως προς τις παρατηρήσεις και όχι ως προς τις επεξηγηματικές μεταβλητές. Παρότι παρουσιάζουν καλή ικανότητα γενίκευσης και υπολογιστική αποτελεσματικότητα, εξακολουθούν να εξαρτώνται από το σύνολο των επεξηγηματικών μεταβλητών, γεγονός που περιορίζει την ερμηνευσιμότητα των μοντέλων. Στη συνέχεια, οι Khan & Zubek (2008) πρότειναν ένα νέο αλγόριθμο SVR

για λογοκριμένα δεδομένα, ο οποίος βασίζεται σε μία τροποποιημένη ασύμμετρη ποινικοποιημένη συνάρτηση απώλειας και σε περιθώρια σφάλματος, με δυνατότητα εφαρμογής τόσο σε δεξιά όσο και σε αριστερά λογοκριμένα δεδομένα. Η σύγκριση της προτεινόμενης μεθόδου με το κλασικό μοντέλο Cox έδειξε σημαντική βελτίωση τόσο στην ακρίβεια των προβλέψεων όσο και στην ικανότητα διάκρισης πληθυσμών ασθενών υψηλού και χαμηλού κινδύνου. Μια εναλλακτική προσέγγιση των SVM προτάθηκε από τους Van Belle et al. (2009a) οι οποίοι ανέπτυξαν μια νέα τεχνική μοντελοποίησης επιβίωσης βασισμένη στις μηχανές υποστήριξης διανυσμάτων ελαχίστων τετραγώνων (Least Squares Support Vector Machine – LS-SVM), συνδυάζοντας τεχνικές κατάταξης (ranking) και παλινδρόμησης. Η συγκεκριμένη προσέγγιση παρουσιάζει ορισμένα σημαντικά πλεονεκτήματα, καθώς η διατύπωση του προβλήματος είναι κυρτή και συνεπώς επιλύεται εύκολα μέσω ενός γραμμικού συστήματος. Παράλληλα, η μη γραμμικότητα εισάγεται μέσω της χρήσης συναρτήσεων πυρήνα ανά συνιστώσα, γεγονός που ενισχύει την ερμηνευσιμότητα του μοντέλου, ενώ η εισαγωγή περιορισμών κατάταξης συμβάλλει στην αποτελεσματικότερη διαχείριση των λογοκριμένων δεδομένων.

Σε πολλές εφαρμογές της Ανάλυσης Επιβίωσης, τα κλινικά δεδομένα χαρακτηρίζονται από μεγάλο αριθμό επεξηγηματικών μεταβλητών σε συνδυασμό με σχετικά μικρό αριθμό παρατηρήσεων, γεγονός που μπορεί να οδηγήσει σε υπολογιστικά και στατιστικά προβλήματα, όπως η υπερπροσαρμογή. Για την αντιμετώπιση των προβλημάτων αυτών, οι Van Belle et al. (2009b) πρότειναν μία τεχνική μείωσης διαστάσεων (feature screening), η οποία βασίζεται στην ανάλυση μέγιστης διακύμανσης για την επιλογή των σημαντικότερων επεξηγηματικών μεταβλητών στο πλαίσιο των LS-SVM για την Ανάλυση Επιβίωσης που είχαν παρουσιάσει στην προηγούμενη εργασία τους (Van Belle et al. (2009a)). Οι βασικές αρχές των μηχανών υποστήριξης διανυσμάτων ελαχίστων τετραγώνων είχαν διατυπωθεί από τους Suykens & Vandewalle (1999) και Suykens et al. (2002). Η προτεινόμενη μέθοδος διαφοροποιείται από τις κλασικές διαδικασίες επιλογής μεταβλητών προς τα εμπρός (forward selection) και προς τα πίσω (backward elimination), καθώς επιλέγει ένα υποσύνολο μεταβλητών ελαχιστοποιώντας τη μέγιστη διακύμανση κάθε μεταβλητής μέσω μιας τροποποιημένης συνάρτησης απώλειας, οδηγώντας τελικά στην ανάπτυξη ενός αραιού μοντέλου πρόβλεψης. Η σύγκριση της συγκεκριμένης μεθόδου με το μοντέλο Cox – Lasso και με το γραμμικό προγνωστικό μοντέλο κλινικής πρακτικής Nottingham Prognostic Index (NPI), τόσο σε πραγματικά όσο και σε προσομοιωμένα δεδομένα, έδειξε ότι παρουσιάζει καλύτερη προγνωστική απόδοση, ενώ παράλληλα επιλέγει αποτελεσματικά τις πραγματικά σημαντικές μεταβλητές που σχετίζονται με το χρόνο επιβίωσης. Ωστόσο, οι ίδιοι επισημαίνουν ότι απαιτείται περαιτέρω διερεύνηση της αποτελεσματικότητας της προτεινόμενης μεθόδου. Στη συνέχεια, οι Van Belle et al. (2011) πρότειναν ένα νέο μοντέλο, το οποίο συνδυάζει την κατάταξη με την παλινδρόμηση Cox, ενσωματώνοντας μετασχηματισμένα μοντέλα στον αλγόριθμο. Με τον τρόπο αυτό επιδίωξαν να συνδέσουν θεωρητικά τη νέα προσέγγιση με την Ανάλυση Επιβίωσης. Ωστόσο, τα αποτελέσματα των προσομοιώσεων έδειξαν ότι το μοντέλο κατάταξης παρουσίασε τη χαμηλότερη απόδοση, ακολούθησε η προτεινόμενη συνδυαστική

προσέγγιση, ενώ η προσέγγιση που βασίζεται αποκλειστικά στην παλινδρόμηση εμφάνισε την καλύτερη προγνωστική απόδοση.

Μία διαφορετική κατεύθυνση για την αντιμετώπιση δεδομένων επιβίωσης υψηλής διάστασης, πέρα από τη μείωση της διάστασης, αποτελεί η ανάπτυξη ενισχυμένων αλγορίθμων που μπορούν να διαχειριστούν αποτελεσματικότερα το μεγάλο όγκο των διαθέσιμων δεδομένων. Στο πλαίσιο αυτό, οι Pölsterl et al. (2015) πρότειναν αλγορίθμους εκπαίδευσης για τρία διαφορετικά μοντέλα SVM, τα οποία βασίζονται στην κατάταξη, στην παλινδρόμηση και στο συνδυασμό των δύο προσεγγίσεων, με στόχο τη βελτιστοποίηση των υφιστάμενων διαδικασιών. Οι προτεινόμενες μέθοδοι ονομάστηκαν Survival Support Vector Machine (SSVM) και βασίζονται στη χρήση της περικομμένης βελτιστοποίησης Newton, καθώς και στην εισαγωγή των στατιστικών δέντρων τάξης (order statistic trees) για τη μείωση της υπολογιστικής πολυπλοκότητας. Η εφαρμογή των συγκεκριμένων τεχνικών πραγματοποιήθηκε στο πλαίσιο γραμμικών μοντέλων. Τα αποτελέσματα έδειξαν ότι οι προτεινόμενες μέθοδοι υπερτερούν των υφιστάμενων διαδικασιών ως προς την απόδοση και την εφαρμογή τους σε προβλήματα γραμμικής χωρικής πολυπλοκότητας. Στη συνέχεια, οι Pölsterl et al. (2016) επέκτειναν το παραπάνω σχήμα βελτιστοποίησης στην περίπτωση μη γραμμικών μοντέλων, καθώς και σε σύνολα δεδομένων με υπολογιστική πολυπλοκότητα της τάξης $O(n^4)$ και $O(pn^6)$, όπου n αποτελεί τον αριθμό του δείγματος και p τον αριθμό των επεξηγηματικών μεταβλητών. Τα αποτελέσματα των προσομοιώσεων και των εφαρμογών σε πραγματικά δεδομένα έδειξαν ότι, σε περιπτώσεις υψηλής δεξιάς λογοκρισίας, όπου το ποσοστό των λογοκριμένων παρατηρήσεων υπερβαίνει το 85%, η προτεινόμενη μέθοδος παρουσιάζει καλύτερη απόδοση σε σύγκριση με τις κλασικές τεχνικές SSVM. Αντίθετα, σε περιπτώσεις με χαμηλότερα ποσοστά λογοκρισίας, οι μέθοδοι εμφανίζουν παρόμοια αποτελέσματα.

3.5.1 Σταθμισμένη Survival LS – SVM τεχνική

Οι Rahmawati et al. (2026) επέκτειναν την εργασία των Van Belle et al. (2009a) σχετικά με την τεχνική LS – SVM, προτείνοντας την εισαγωγή σταθμισμένων βαρών στις λογοκριμένες και μη λογοκριμένες παρατηρήσεις μέσω της μεθόδου Weighted Least Squares Support Vector Machine (WLSSVM). Η ιδέα της προτεινόμενης μεθόδου βασίζεται στο γεγονός ότι ολοένα και περισσότερα σύνολα δεδομένων χαρακτηρίζονται από υψηλά ποσοστά λογοκρισίας. Η εφαρμογή των κλασικών τεχνικών σε τέτοιες περιπτώσεις μπορεί να οδηγήσει στην εισαγωγή μεροληπτικής πληροφορίας και, κατά συνέπεια, σε μη ακριβή αποτελέσματα.

Η συγκεκριμένη τεχνική ενσωματώνει την εκτίμηση της s.f. μέσω του κλασικού εκτιμητή Kaplan – Meier, η οποία ορίζεται με τον ακόλουθο συμβολισμό:

$$\hat{S}(t) = \prod_{t_i \leq t} \left(\frac{N_{i-1} - d_i}{N_{i-1}} \right)$$

όπου N_{i-1}: αποτελεί τον αριθμό των παρατηρήσεων που βρίσκονται σε κίνδυνο τη χρονική στιγμή t_i, ενώ d_i: αποτελεί τον αριθμό των γεγονότων που παρατηρούνται τη χρονική στιγμή t_i.

Η κλασική μορφή του LS – SVM δεν ενσωματώνει βάρη στη διαδικασία εκπαίδευσης του μοντέλου, γεγονός που μπορεί να την καταστήσει ευαίσθητη σε ακραίες τιμές ή σε μη αντιπροσωπευτικά δεδομένα. Για τον λόγο αυτό, προτείνεται η ελαχιστοποίηση μιας σταθμισμένης συνάρτησης απώλειας, η οποία περιλαμβάνει τόσο τον όρο σφάλματος όσο και τον όρο κανονικοποίησης για τον ορισμό της πολυπλοκότητας του μοντέλου. Η συγκεκριμένη συνάρτηση απώλειας διατυπώνεται ως:

$$\min_{w,\xi} \left[\frac{1}{2} \boldsymbol{w}^T \boldsymbol{w} + \frac{1}{2} C \sum\nolimits_{i=1}^{n} \sum\nolimits_{j=1}^{n} r_{ij} B_{ij} \xi_{ij}^2 \right]$$

υπό την προϋπόθεση ότι:

$$\boldsymbol{w}^T \varphi(\boldsymbol{x}_j) - \boldsymbol{w}^T \varphi(\boldsymbol{x}_i) = 1 - \xi_{ij}, \qquad \forall i, j = 1, \dots, n \;\; \text{όπου}$$

- $\varphi(\cdot)$: συνάρτηση χαρτογράφησης (mapping function) σε έναν χώρο χαρακτηριστικών υψηλής διάστασης.
- C: παράμετρος κανονικοποίησης.
- ξ_{ij}: μεταβλητή χαλάρωσης (slack variable).
- B_{ji}: αντιπροσωπεύει το βάρος για το i – οστό και j – οστό άτομο που μπορεί να συγκριθεί.
- $r_{ij} = \begin{cases} 1, (t_i < t_j, \delta_i = 1) \\ 0, (t_i \geq t_j, \delta_i = 0) \end{cases}$: δείκτης σύγκρισης μεταξύ του i – οστού και j – οστού ατόμου.

Για τη βελτιστοποίηση των παραμέτρων του μοντέλου κατασκευάζεται η ακόλουθη Lagrangian συνάρτηση:

$$L(\boldsymbol{w}, \xi, \alpha) = \frac{1}{2} \boldsymbol{w}^T \boldsymbol{w} + \frac{1}{2} C \sum\nolimits_{i=1}^{n} \sum\nolimits_{i<j}^{n} r_{ij} B_{ij} \xi_{ij}^2 - \sum\nolimits_{i=1}^{n} \sum\nolimits_{j=1}^{n} \boldsymbol{a}_{ij} \left(\boldsymbol{w}^T \varphi(\boldsymbol{x}_j) - \boldsymbol{w}^T \varphi(\boldsymbol{x}_i) - 1 + \xi_{ij} \right)$$

όπου $\boldsymbol{a}_{ij}$: είναι οι Lagrangian πολλαπλασιαστές. Με την εφαρμογή της διαδικασίας ελαχιστοποίησης της παραπάνω συνάρτησης προκύπτει ότι ο προγνωστικός δείκτης εκτιμάται μέσω του ακόλουθου τύπου:

$$\hat{u}(x^*) = \hat{\boldsymbol{a}}^T \boldsymbol{D} \boldsymbol{K}(x^*)$$

- $\hat{\boldsymbol{\alpha}}_{(m \times 1)} = (C \boldsymbol{B} \boldsymbol{D} \boldsymbol{K} \boldsymbol{D}^T + \boldsymbol{I})^{-1} C \boldsymbol{B} \boldsymbol{1}_{(m \times 1)}$: η εκτίμηση των Lagrangian πολλαπλασιαστών
- $\boldsymbol{D}_{(m \times n)}$: ένας πίνακας με στοιχεία $\{-1,0,1\}$, ώστε να ισχύει η εξής σχέση:

$$\boldsymbol{D}\boldsymbol{X} = \begin{bmatrix} x_1 - x_2 \\ \vdots \\ x_1 - x_n \\ \vdots \\ x_{n-1} - x_n \end{bmatrix} \;\; \text{όπου} \;\; \boldsymbol{X} = (x_1, \dots, x_n)^T$$

- **K**: είναι ένας πίνακας $n \times n$ διαστάσεων με στοιχείο $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\left|\left|x_i - x_j\right|\right|^2}{2\sigma^2}\right)$, ο οποίος περιλαμβάνει τη συνάρτηση πυρήνα μεταξύ του i – οστού και j – οστού ατόμου.
- $\boldsymbol{I}_{(m\times m)} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$**:** ο μοναδιαίος πίνακας διαστάσεων $m \times m$.
- $\boldsymbol{B} = \begin{bmatrix} B_{12} & 0 & & 0 & \cdots & 0 \\ 0 & B_{13} & & 0 & \cdots & 0 \\ 0 & 0 & B_{14} & \cdots & & 0 \\ \vdots & \vdots & \vdots & \ddots & & \vdots \\ 0 & 0 & 0 & \cdots & B_{(m-1)m} & \end{bmatrix}$: αποτελεί τον πίνακα των βαρών.

Για τις λογοκριμένες παρατηρήσεις η τιμή B_{ij} είναι μία σταθερά, ενώ στην περίπτωση εμφάνισης συμβάντος λαμβάνει την τιμή $B_{ij} = S(t_j)$. Η εκτίμηση των υπόλοιπων παραμέτρων που συμμετέχουν στη συνάρτηση απώλειας πραγματοποιείται μέσω του αλγορίθμου Particle Swarm Optimization (PSO). Η τελική επιλογή των συγκεκριμένων παραμέτρων μπορεί να επηρεάσει τόσο θετικά όσο και αρνητικά την τελική απόδοση του μοντέλου.

Η προτεινόμενη τεχνική αποτελεί μία υπολογιστικά αποτελεσματική μέθοδο, η οποία αντιμετωπίζει βασικά προβλήματα των δεδομένων, όπως ο θόρυβος και η παρουσία ακραίων τιμών, ενώ παράλληλα παρουσιάζει χαμηλό υπολογιστικό κόστος και υψηλή ικανότητα γενίκευσης. Σύμφωνα με τα αποτελέσματα των προσομοιώσεων, η μέθοδος παρουσιάζει καλύτερη απόδοση σε σύγκριση με τις κλασικές τεχνικές που εξετάστηκαν καθώς και υψηλή ακρίβεια πρόβλεψης. Η ενσωμάτωση της συνάρτησης πυρήνα παρέχει στη μεθοδολογία τη δυνατότητα ανίχνευσης μη γραμμικών προτύπων, τα οποία δεν είναι εύκολα ανιχνεύσιμα μέσω του κλασικού μοντέλου Cox. Επιπλέον, η μέθοδος παρουσιάζει ικανοποιητική απόδοση τόσο σε μικρά όσο και σε μεγάλα σύνολα δεδομένων. Ωστόσο, ένα σημαντικό μειονέκτημα της συγκεκριμένης προσέγγισης αποτελεί η ευαισθησία ως προς την επιλογή βαρών που μπορεί να προκληθεί εξαιτίας της χρήσης του εκτιμητή Kaplan – Meier. Για αυτό το λόγο, οι Rahmawati et al. (2026) προτείνουν ως μελλοντική κατεύθυνση τη διερεύνηση της χρήσης ενός Bayesian Kaplan – Meier εκτιμητή για τη στάθμιση των λογοκριμένων παρατηρήσεων ή την εφαρμογή τεχνικών στάθμισης που βασίζονται στην IPCW.

4ο Κεφάλαιο
Εφαρμογή και Σύγκριση Τεχνικών σε πραγματικό σύνολο δεδομένων επιβίωσης υψηλής διαστατικότητας

4.1 Εισαγωγή

Στόχος των προηγούμενων κεφαλαίων ήταν να κατανοήσει ο αναγνώστης τη χρησιμότητα και την αναγκαιότητα των τεχνικών φιλτραρίσματος (feature screening) σε δεδομένα επιβίωσης υψηλής διάστασης, καθώς και την εφαρμογή μεθόδων στατιστικής ποινικοποίησης (regularization) για τον εντοπισμό των πλέον επιδραστικών επεξηγηματικών μεταβλητών. Επιπλέον, επιδιώχθηκε η απόκτηση μιας ευρείας εικόνας σχετικά με τις μεθοδολογίες που έχουν αναπτυχθεί βιβλιογραφικά έως σήμερα, με ιδιαίτερη έμφαση στη σύγχρονη εποχή, όπου οι διαθέσιμοι υπολογιστικοί πόροι και οι αλγόριθμοι έχουν εξελιχθεί ραγδαία. Τέλος, παρουσιάστηκαν σύγχρονες τεχνικές μηχανικής μάθησης, οι οποίες επιτρέπουν στον αναλυτή την ανάπτυξη ισχυρών και αξιόπιστων προγνωστικών μοντέλων επιβίωσης. Εφοδιασμένος με αυτά τα μεθοδολογικά εργαλεία, ο αναλυτής είναι πλέον σε θέση να αναζητήσει το βέλτιστο προγνωστικό μοντέλο ανάλογα με το εκάστοτε ερευνητικό ερώτημα, επιτυγχάνοντας δραστική μείωση της διάστασης των δεδομένων, διατηρώντας παράλληλα το μέγιστο δυνατό ποσοστό της αρχικής πληροφορίας.

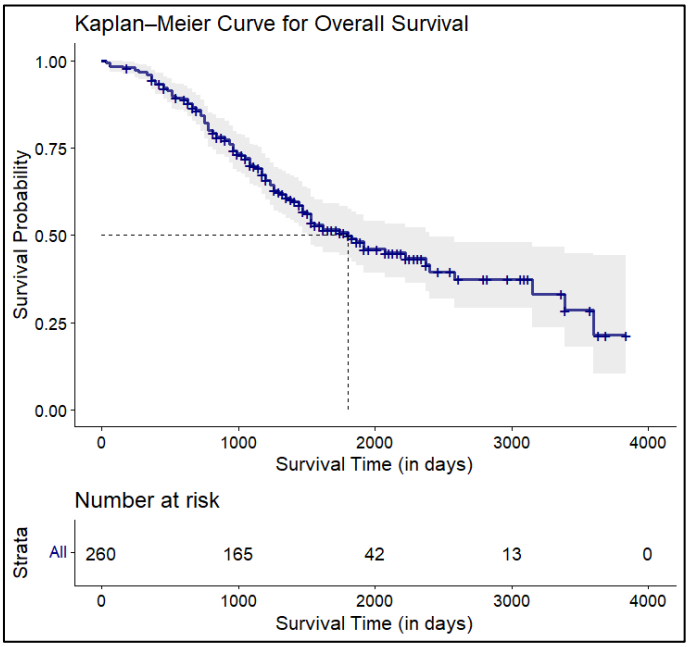
Το σύνολο δεδομένων το οποίο θα αναλυθεί στη συνέχεια αφορά γονιδιωματικά και κλινικά δεδομένα ασθενών με προχωρημένου σταδίου ορώδη καρκίνο των ωοθηκών υψηλού βαθμού κακοήθειας (advanced – stage high – grade serous ovarian cancer). Ο συγκεκριμένος τύπος καρκίνου αποτελεί τον πιο επιθετικό ιστολογικό τύπο της νόσου, ο οποίος συνήθως διαγιγνώσκεται σε προχωρημένο στάδιο λόγω της απουσίας πρώιμων συμπτωμάτων και αξιόπιστων διαγνωστικών δοκιμασιών. Τα δεδομένα αυτά αντλήθηκαν από τη διεθνή βάση δεδομένων Gene Expression Omnibus (GEO) με τους κωδικούς πρόσβασης GSE32062 και GSE32063. Το σύνολο δεδομένων GSE32062, το οποίο θα χρησιμοποιηθεί ως το κύριο σύνολο εκπαίδευσης (training set) για την ανάπτυξη των μοντέλων εφαρμογής, περιλαμβάνει 260 ασθενείς ιαπωνικής καταγωγής που διαγνώστηκαν μεταξύ Ιουλίου 1997 και Ιουνίου 2010. Σε κάθε ασθενή καταγράφηκε το κλινικό στάδιο της νόσου (“Tumor Grade”) σύμφωνα με τα κριτήρια της FIGO (Στάδιο III ή IV) με το στάδιο III να περιλαμβάνει τα υποστάδια IIIa, IIIb, IIIc, ο βαθμός ιστολογικής διαφοροποίησης του όγκου (“Grade”) με βάση την ταξινόμηση Silverberg (διαχωρισμένος σε “Grade 3” = “high” & “Grade 2” = “low”) καθώς και η χειρουργική κατάσταση της κυτταρομείωσης (“Surgery” – debulking status) μετά την αρχική επέμβαση. Η χειρουργική κατάσταση κατηγοριοποιήθηκε ως βέλτιστη (optimal surgery) όταν η υπολειμματική νόσος μετά το χειρουργείο ήταν μικρότερη από 1cm, ή ως μη βέλτιστη (suboptimal surgery) όταν η υπολειμματική νόσος παρέμενε ίση ή μεγαλύτερη από 1cm. Επιπλέον, καταγράφηκε ο χρόνος συνολικής επιβίωσης (overall survival time), ο οποίος υπολογίστηκε ως το μεσοδιάστημα από την ημερομηνία της αρχικής χειρουργικής

επέμβασης έως το θάνατο εξαιτίας του καρκίνου των ωοθηκών, καθώς και η κατάσταση επιβίωσης. Όλες οι ασθενείς έλαβαν την καθιερωμένη μετεγχειρητική χημειοθεραπεία πρώτης γραμμής με βάση την πλατίνα και τις ταξάνες. Τα δείγματα RNA απομονώθηκαν από φρέσκο – κατεψυγμένο (fresh – frozen) ιστό του πρωτοπαθούς όγκου κατά τη διάρκεια του χειρουργείου, εφόσον η ιστολογική αξιολόγηση επιβεβαίωσε την παρουσία καρκινικών κυττάρων σε ποσοστό άνω του 70%. Η μέτρηση της γονιδιακής έκφρασης (συμπεριλαμβανομένων 20.106 γονιδίων) πραγματοποιήθηκε με τη χρήση της πλατφόρμας μικροσυστοιχίων (microarrays) Agilent Whole Human Genome Oligo Microarray. Στο συγκεκριμένο σύνολο δεδομένων εκπαίδευσης, το πλήθος των δεξιά λογοκριμένων (right – censored) παρατηρήσεων ανέρχεται σε 139 ασθενείς (ποσοστό 53,46%) έναντι 121 ασθενών όπου είναι διαθέσιμοι οι πλήρεις χρόνοι επιβίωσης (συμβάντα θανάτου). (Yoshihara et al. (2012))

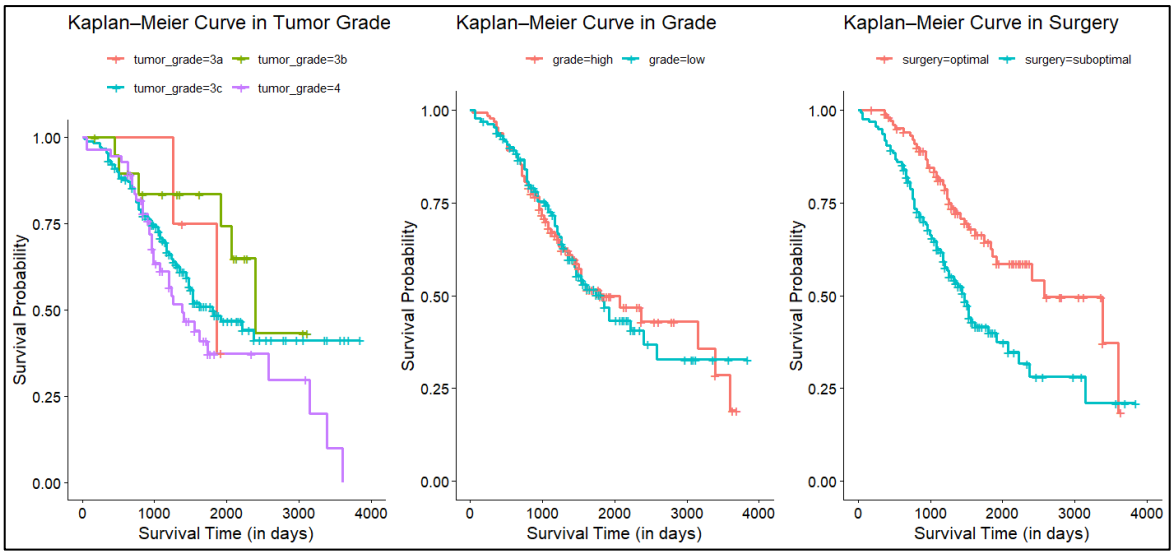
Παρακάτω παρουσιάζονται οι πίνακες συχνότητας των κλινικών μεταβλητών καθώς και οι καμπύλες Kaplan – Meier στις οποίες φαίνεται πως κυμαίνεται ο χρόνος επιβίωσης των ασθενών στα διάφορα επίπεδα των κλινικών μεταβλητών που αναφέρθηκαν πρωτύτερα.

tumor_grade	**grade**	**surgery**
3a: 4	high:129	optimal :103
3b: 20	low :131	suboptimal:157
3c:180		
4 : 56		

Πίνακας 4. 1 Πίνακας Συχνοτήτων κλινικών μεταβλητών του GSE32062



Εικόνα 4. 1 Καμπύλη Kaplan - Meier για το συνολικό χρόνο επιβίωσης



Εικόνα 4. 2 Καμπύλες Kaplan - Meier του χρόνου επιβίωσης ανά κλινική μεταβλητή

Από την Εικόνα 4. 2 φαίνεται ότι για τις μεταβλητές "Tumor Grade" & "Grade", η υπόθεση αναλογικού κινδύνου παραβιάζεται εν αντιθέσει με τη μεταβλητή "Surgery" που υπάρχει μόνο ένα σημείο διασταύρωσης σε μεγάλο χρονικό σημείο επιβίωσης. Για τον περιορισμό του θορύβου και την αφαίρεση γονιδίων με χαμηλή πληροφορία, εφαρμόστηκε φιλτράρισμα με βάση τη Διάμεση Απόλυτη Απόκλιση (Median Absolute Deviation – MAD), από όπου επιλέχθηκαν τα 3.000 γονίδια με τη robust διακύμανση. Η χρήση της MAD ως δείκτη εύρωστης διακύμανσης κρίνεται επιβεβλημένη έναντι της κλασικής διακύμανσης ή τυπικής απόκλισης, καθώς βασίζεται στη διάμεσο και ως εκ τούτου παραμένει ανθεκτική στην παρουσία ακραίων τιμών (outliers) και ασύμμετρων κατανομών. Με τον τρόπο αυτό, διασφαλίζεται βιολογική διαφοροποίηση, θωρακίζοντας το μοντέλο από τεχνικά σφάλματα ή μεμονωμένες ακραίες μετρήσεις των μικροσυστοιχιών που θα αλλοίωναν τεχνητά τις παραδοσιακές παραμετρικές εκτιμήσεις διασποράς. Τα δεδομένα της γονιδιακής έκφρασης ακολούθως τυποποιήθηκαν (Z – score normalization) ώστε να έχουν μέσο όρο 0 και διακύμανση 1.

4.2 Δείκτης Συμφωνίας του Harrell

Για την αξιολόγηση και τη σύγκριση της προγνωστικής ικανότητας των αναπτυχθέντων μοντέλων, χρησιμοποιείται ο Δείκτης Συμφωνίας (Concordance Index / C–Index), ο οποίος αποτελεί ένα εύκολα ερμηνεύσιμο μέτρο της προγνωστικής διακριτικής ικανότητας (predictive discrimination) ενός μοντέλου, ιδιαίτερα σε δεδομένα επιβίωσης με παρουσία λογοκρισίας. Όπως περιγράφεται στη κλασική εργασία των Harrell et al. (1996), ο δείκτης *c* είναι ένα ευρέως εφαρμόσιμο μέτρο που μπορεί να χρησιμοποιηθεί σε συνεχή, διχοτομικά, διατάξιμα, καθώς και σε λογοκριμένα δεδομένα χρόνου μέχρι το συμβάν. Σχετίζεται άμεσα με τη συσχέτιση κατάταξης (rank correlation) μεταξύ των προβλεπόμενων και των παρατηρούμενων αποτελεσμάτων, αποτελώντας μια τροποποίηση του δείκτη συσχέτισης κατάταξης τύπου Kendall – Goodman – Kruskal – Somers.

Μαθηματικά, ο δείκτης c ορίζεται ως η αναλογία όλων των αξιοποιήσιμων ζευγών ασθενών στα οποία οι προβλέψεις του μοντέλου και η πραγματική έκβαση είναι συμβατές. Στην περίπτωση της πρόβλεψης του χρόνου μέχρι το θάνατο, ο υπολογισμός του δείκτη γίνεται εξετάζοντας όλα τα δυνατά ζευγάρια ασθενών, όπου τουλάχιστον ένας από τους δύο έχει καταλήξει. Οι προβλέψεις για ένα ζεύγος θεωρούνται συμβατές με την πραγματική έκβαση εάν ο προβλεπόμενος χρόνος επιβίωσης (ή η προβλεπόμενη πιθανότητα επιβίωσης έως ένα χρονικό σημείο) είναι μεγαλύτερος για τον ασθενή που έζησε πραγματικά περισσότερο. Αν ένας ασθενής έχει πεθάνει και ο άλλος είναι γνωστό ότι επιβίωσε τουλάχιστον μέχρι το χρόνο θανάτου του πρώτου, θεωρείται ότι ο δεύτερος έζησε περισσότερο από τον πρώτο. Όταν οι προβλεπόμενες τιμές επιβίωσης είναι πανομοιότυπες για ένα ζεύγος ασθενών (ισοβαθμία πρόβλεψης), προστίθεται η τιμή 0,5 στον αριθμητή του δείκτη c, ενώ το ζεύγος εξακολουθεί να προσμετράται κανονικά στον παρονομαστή ως αξιοποιήσιμο. Αντίθετα, ένα ζεύγος ασθενών χαρακτηρίζεται ως μη αξιοποιήσιμο και αποκλείεται από την ανάλυση (τόσο από τον αριθμητή όσο και από τον παρονομαστή) όταν:

1) Και οι δύο ασθενείς πέθαναν ακριβώς την ίδια χρονική στιγμή
2) Ο ένας ασθενής πέθανε, αλλά ο άλλος είναι ακόμη ζωντανός (λογοκριμένος) και δεν έχει παρακολουθηθεί για αρκετό διάστημα ώστε να διαπιστωθεί αν έζησε περισσότερο από αυτόν που απεβίωσε.

Αντί του προβλεπόμενου χρόνου επιβίωσης, για τον υπολογισμό του c μπορεί να χρησιμοποιηθεί ισοδύναμα η προβλεπόμενη πιθανότητα επιβίωσης έως ένα σταθερό χρονικό σημείο, υπό την προϋπόθεση ότι οι δύο εκτιμήσεις συνδέονται με αμφιμονοσήμαντη σχέση, (one – to – one function), κάτι που ισχύει όταν ικανοποιείται η υπόθεση αναλογικού κινδύνου.

Ο δείκτης c εκτιμά την πιθανότητα συμφωνίας των προβλεπόμενων και των παρατηρούμενων εκβάσεων, λαμβάνοντας τιμές στο διάστημα [0, 1] με την ακόλουθη ερμηνεία:

- Μία τιμή ίση με 0 αντιστοιχεί σε τέλειο αντίστροφο προγνωστικό διαχωρισμό των ασθενών
- Μία τιμή ίση με 0,5, υποδηλώνει πλήρη απουσία προγνωστικής διακριτικής ικανότητας (ισοδύναμη με τυχαία επιλογή)
- Μία τιμή ίση με 1 αντιστοιχεί σε τέλειο προγνωστικό διαχωρισμό των ασθενών
- Για τους αναλυτές που προτιμούν έναν συντελεστή συσχέτισης κατάταξης που κυμαίνεται από -1 έως +1 (όπου το μηδέν (0) σημαίνει απουσία συσχέτισης), μπορεί να υπολογιστεί ο δείκτης συσχέτισης κατάταξης Somers' D μέσω της μαθηματικής σχέσης:

$$D_{xy} = 2(c - 0{,}5)$$

Τέλος, αν και οι δείκτες κατάταξης όπως ο c είναι ευρέως εφαρμόσιμοι και εύκολα ερμηνεύσιμοι, παρουσιάζουν το μειονέκτημα ότι δεν είναι ιδιαίτερα ευαίσθητοι στον εντοπισμό μικρών διαφορών στη διακριτική ικανότητα μεταξύ δύο μοντέλων. Αυτό συμβαίνει επειδή η μέθοδος κατάταξης αντιμετωπίζει όλα τα συμβατά ζεύγη με τον

ίδιο τρόπο, ανεξάρτητα από το μέγεθος της διαφοράς στις προβλεπόμενες πιθανότητες (για παράδειγμα, θεωρεί ένα ζεύγος με προβλέψεις κινδύνου (0,01) και (0,90) ως εξίσου συμβατό με ένα ζεύγος που έχει προβλέψεις (0,05) και (0,80).

4.3 Φιλτράρισμα Μεταβλητών Υψηλής Διάστασης μέσω RMSTSIS και Διαδικασία Σταθερής Επιλογής Μεταβλητών

Λόγω της εξαιρετικά υψηλής διάστασης του γονιδιωματικού συνόλου δεδομένων (3.000 γονίδια έναντι μόλις 260 ασθενών στο training set), η απευθείας εφαρμογή σύνθετων αλγόριθμων μηχανικής μάθησης ενέχει σοβαρό κίνδυνο υπερπροσαρμογής (overfitting). Για την αντιμετώπιση της «κατάρας της διάστασης», εφαρμόστηκε η μέθοδος φιλτραρίσματος RMSTSIS των Chen et al. (2024). Η μέθοδος αυτή είναι model – free, δεν απαιτεί την υπόθεση αναλογικού κινδύνου και επιτρέπει τον εντοπισμό τόσο γραμμικών όσο και μη γραμμικών συσχετίσεων με το χρόνο επιβίωσης, χρησιμοποιώντας ως μέτρο τον περιορισμένο μέσο χρόνο επιβίωσης (RMST) αντί για την κλασική συνάρτηση κινδύνου.

Το θεωρητικό κατώφλι για τον αριθμό επιλεγμένων γονιδίων κατά το screening ορίστηκε με βάση τη μαθηματική σχέση:

$$p_{train} = \lfloor n_{train} / \log(n_{train}) \rfloor$$

Για το σύνολο εκπαίδευσης των $n_{train} = 260$ ασθενών, το κατώφλι αυτό απέδωσε ακριβώς $p_{train} = 47$ γονίδια. Για να διασφαλιστεί η σταθερότητα και η αξιοπιστία των επιλεγμένων βιοδεικτών, ενσωματώθηκε μια επαναληπτική διαδικασία επιλογής σταθερότητας (stability selection) με 25 επαναλαμβανόμενους τυχαίους διαχωρισμούς (repeated random splits). Σε κάθε μία από τις 25 επαναλήψεις:

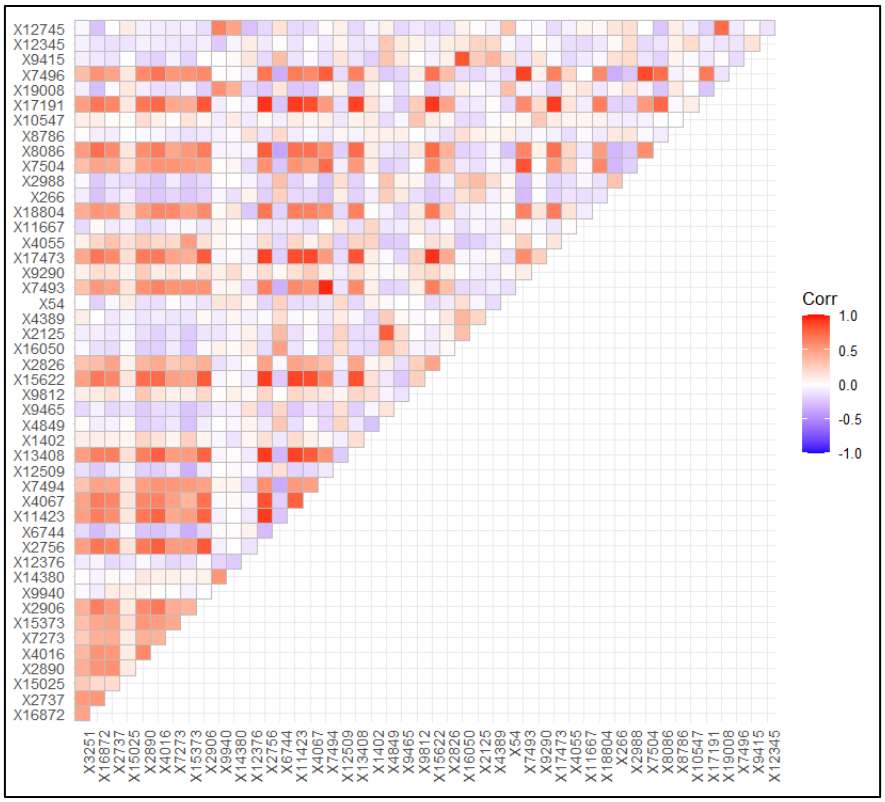
1) Τα δεδομένα διαχωρίζονται τυχαία 80% train ($n = 208$) και 20% validation set.
2) Εφαρμόζονται ο αλγόριθμος RMSTSIS στο train set για την επιλογή των 44 σημαντικότερων γονιδίων
3) Αποθηκεύονται τα $p = n/\log(n) = 39$ σημαντικά γονίδια με βάση το RMSTSIS.

Τα αναλυτικά αποτελέσματα των 25 επαναλήψεων παρατίθενται στον Πίνακα 4.9.

Με την ολοκλήρωση των 25 επαναλήψεων, καταγράφηκε η συχνότητα επιλογής (frequency) κάθε γονιδίου. Τα 47 γονίδια με την υψηλότερη συχνότητα εμφάνισης επιλέχθηκαν για να αποτελέσουν το τελικό σταθεροποιημένο σύνολο μεταβλητών. Στην κορυφή της λίστας αυτής βρέθηκαν γονίδια με εξαιρετικά υψηλή σταθερότητα, όπως το Gene 3251 (εμφανίστηκε στις 24 επαναλήψεις), το Gene 16872 (23 εμφανίσεις) και το Gene 2737 (22 εμφανίσεις). Η συσχέτιση μεταξύ αυτών των 47 κορυφαίων γονιδίων αναλύθηκε με Spearman correlation, αποκαλύπτοντας έντονες συνεκφράσεις και βιολογικά μονοπάτια (Εικόνα 4. 3). Η λίστα με τα IDs των κορυφαίων 47 γονιδίων παρατίθεται στον Πίνακα 4.2.

X3251	X7273	X2756	X13408	X2826	X9290	X2988	X19008
X16872	X15373	X6744	X1402	X16050	X17473	X7504	X7496
X2737	X2906	X11423	X4849	X2125	X4055	X8086	X9415
X15025	X9940	X4067	X9465	X4389	X11667	X8786	X12345
X2890	X14380	X7494	X9812	X54	X18804	X10547	X12745
X4016	X12376	X12509	X15622	X7493	X266	X17191	

Πίνακας 4. 2 IDs των κορυφαίων 47 μεταβλητών με την υψηλότερη συχνότητα



Εικόνα 4. 3 Γράφημα Συσχετίσεων Spearman για τις 47 κορυφαίες μεταβλητές

4.4 Ανάπτυξη και Βελτιστοποίηση Μοντέλου Πλάγιων Τυχαίων Δασών Επιβίωσης

Τα Πλάγια Τυχαία Δάση Επιβίωσης αποτελούν μια σύγχρονη και εξαιρετικά ισχυρή επέκταση των κλασικών δασών επιβίωσης, καθώς πραγματοποιούν διαχωρισμούς κόμβων χρησιμοποιώντας γραμμικούς συνδυασμούς μεταβλητών αντί για ορθογώνιους διαχωρισμούς μιας μόνο μεταβλητής. Στην παρούσα μελέτη, για την εύρεση των βέλτιστων συντελεστών των πλάγιων διαχωρισμών στους εσωτερικούς κόμβους των δέντρων, ενσωματώθηκε η μέθοδος της ποινικοποιημένης παλινδρόμησης Cox με ποινή Elastic Net (Penalized Cox Regression). Η προσέγγιση αυτή επιτρέπει στο μοντέλο να διαχειρίζεται αποτελεσματικά τις συσχετίσεις και τις αλληλεπιδράσεις μεταξύ των 47 σταθερών γονιδίων που προέκυψαν από τη διαδικασία του RMST Screening και των 3 κλινικών μεταβλητών.

4.4.1 Ρύθμιση Υπερπαραμέτρων

Για την εύρεση της βέλτιστης αρχιτεκτονικής του μοντέλου ORSF στο σύνολο εκπαίδευσης, εφαρμόστηκε συστηματική αναζήτηση σε πλέγμα τιμών (Grid Search) με 1000 δέντρα ανά δάσος ($n_{tree} = 1000$). Εκτελέστηκε 5 – fold CV για την ανάδειξη του βέλτιστου συνδυασμού των παραμέτρων με βάση τη μεγιστοποίηση του δείκτη C – Index. Σε κάθε κόμβο, οι μεταβλητές επιλέχθηκαν με βάση τον log – rank έλεγχο του Cox μοντέλου. Οι παράμετροι που βελτιστοποιήθηκαν ήταν ο αριθμός των υποψήφιων μεταβλητών προς εξέταση σε κάθε διαχωρισμό ($mtry = \{5, 10, 15, 20\}$) και η παράμετρος ανάμειξης της ποινής Elastic Net ($alpha = \{0, 0.25, 0.5, 0.75, 1\}$).

Τα αποτελέσματα της διαδικασίας Grid Search παρουσιάζονται αναλυτικά στον Πίνακα 4.3.

mtry	alpha	mean_cindex	sd_cindex	min_cindex	max_cindex
20	1	0,65250	0,01306	0,63510	0,66993
20	0,5	0,65607	0,01470	0,63763	0,67832
15	0,5	0,66351	0,01748	0,63763	0,68671
20	0,25	0,65618	0,01800	0,63763	0,68531
15	1	0,66356	0,01899	0,64646	0,69510
15	0,75	0,66580	0,02038	0,64394	0,69930
15	0,25	0,66627	0,02049	0,64520	0,69930
20	0,75	0,65607	0,02127	0,62753	0,68671
10	0,5	0,67546	0,02489	0,64646	0,71189
10	1	0,66924	0,02545	0,64268	0,70629
10	0,75	0,67192	0,02824	0,64646	0,71888
5	1	0,68361	0,02826	0,65891	0,73007
10	0,25	0,66894	0,02975	0,64015	0,71469
5	0,5	0,68014	0,03263	0,64729	0,73007
5	0,75	0,68799	0,03345	0,65504	0,73706
5	0,25	0,67906	0,03569	0,63953	0,73147
5	0	0,68417	0,03983	0,63307	0,74126
10	0	0,67719	0,04072	0,61370	0,72587
20	0	0,67244	0,04382	0,59690	0,70909
15	0	0,66891	0,04475	0,59432	0,71469

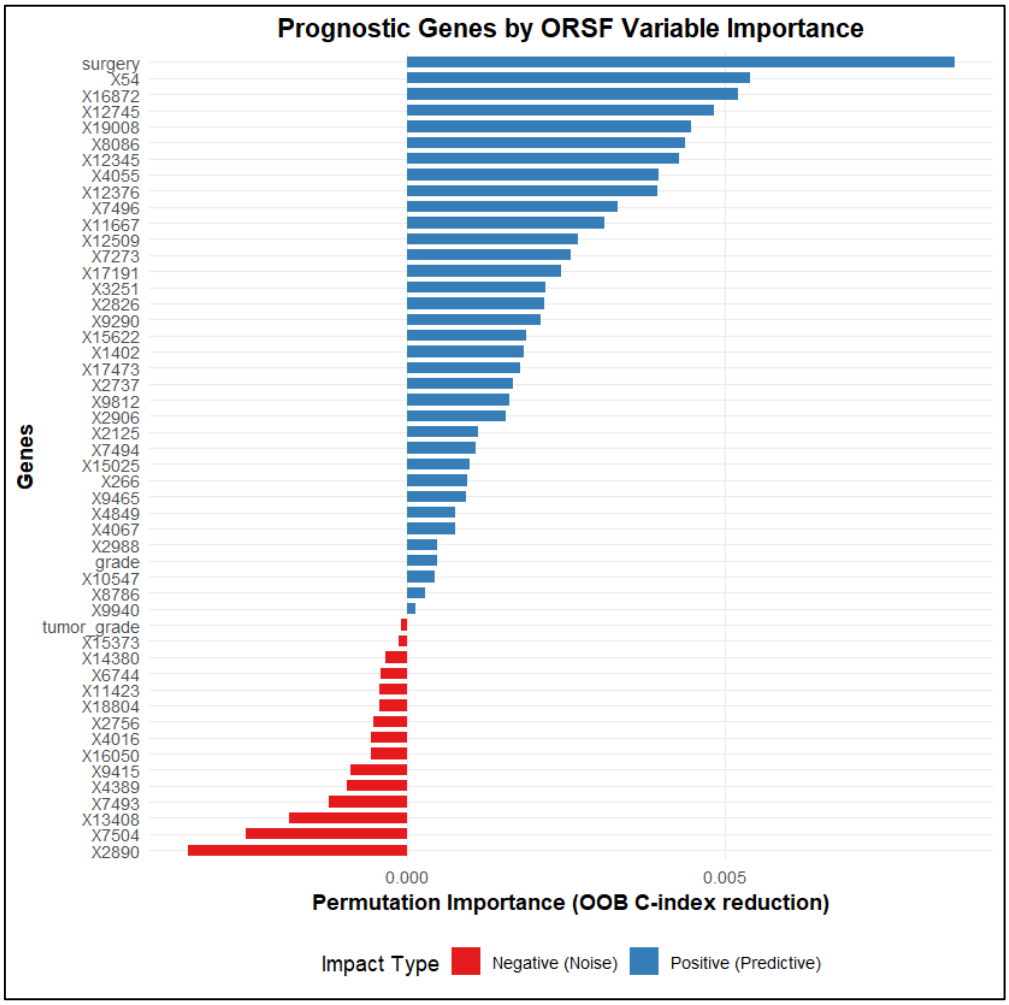
Πίνακας 4. 3 Αποτελέσματα βελτιστοποίησης υπερπαραμέτρων του μοντέλου ORSF (Grid Search)

Οι μέσες τιμές των C – Index που προέκυψαν από την 5 – fold CV διαδικασία είναι αρκετά κοντά και γι' αυτόν τον λόγο δεν θα επιλεγεί ο συνδυασμός με τη μεγαλύτερη μέση τιμή, αλλά με το σταθερότερο αποτέλεσμα. Παρατηρείται ότι ο συνδυασμός $mtry = 20 \; και \; alpha = 1$ έχει τη μικρότερη τυπική απόκλιση, το οποίο σημαίνει ότι σε κάθε κόμβο εξετάζεται ένα πλήθος 20 μεταβλητών για το διαχωρισμό του κόμβου εφαρμόζοντας ένα μοντέλο Cox με ποινή Lasso (L_1). Ωστόσο, για πιο ευέλικτα αποτελέσματα και αποφυγή υπερπροσαρμογής, επιλέγεται ο αμέσως επόμενος

καλύτερος συνδυασμός παραμέτρων $mtry = 20$ $και$ $alpha = 0{,}5$ ο οποίος ισορροπεί μεταξύ των ποινών Ridge (L_2) και Lasso (L_1).

Στην Εικόνα 4. 4 παρουσιάζεται η σχετική σημαντικότητα των προγνωστικών μεταβλητών (Variable Importance – VIMP), όπως αυτή υπολογίστηκε μέσω της μεθόδου της τυχαίας μετάθεσης (Permutation Importance) στο βέλτιστο μοντέλο ORSF. Η συγκεκριμένη μέθοδος αξιολογεί τη συνεισφορά κάθε μεταβλητής ξεχωριστά, προβαίνοντας σε τυχαία αναδιάταξη (permutation) των τιμών της στα Out – of – Bag (OOB) δεδομένα, απαλείφοντας έτσι τη σχέση της με το καταληκτικό σημείο επιβίωσης. Η μείωση που επιφέρει αυτή η διαδικασία στο δείκτη προγνωστικής ακρίβειας (C – Index) αποτελεί το μέτρο σημαντικότητας της μεταβλητής.

- Στην κορυφή της προγνωστικής σημασίας βρέθηκε η κλινική μεταβλητή "Surgery" και ακολουθούν τα γονίδια X54, X16872, X12745.
- Αντίθετα, 14 γονίδια και η κλινική μεταβλητή "Tumor Grade" αναγνωρίστηκαν ως «θόρυβος» (Permutation Impact / Noise), καθώς η παρουσία τους μειώνει ελάχιστα την OOB ακρίβεια του μοντέλου. Τέτοια γονίδια είναι τα X2890, X7504, X13408, X7493.



Εικόνα 4. 4 Σημαντικότητα Μεταβλητών με βάση το πλήρες μοντέλο ORSF

4.4.2 Αναδρομική Απαλοιφή

Με στόχο τη δημιουργία μιας κλινικά εφαρμόσιμης, ερμηνεύσιμης και εξαιρετικά αραιής γονιδιακής υπογραφής, εφαρμόστηκε ο αλγόριθμος αναδρομικής απαλοιφής μεταβλητών (recursive variable elimination – "orsf_vs"). Η διαδικασία αυτή ξεκινά από το πλήρες μοντέλο των 47 γονιδίων και αφαιρεί προοδευτικά σε κάθε βήμα το γονίδιο με τη χαμηλότερη σημαντικότητα, επανεκπαιδεύοντας το δάσος και αξιολογώντας τη μεταβολή της ακρίβειας. Ο Πίνακας 4.10 περιέχει αναλυτικά τα βήματα της διαδικασίας μέχρι τις 10 κορυφαίες μεταβλητές. Η ανάλυση έδειξε ότι η μείωση του αριθμού των γονιδίων από 52 (47 γονίδια και 5 dummy variables που προκύπτουν από τις 3 κλινικές μεταβλητές) σε 10 επιτρέπει τη διατήρηση της προγνωστικής ισχύος σε υψηλά επίπεδα (OOB C – Index = 0,6973 για το sparse μοντέλο έναντι 0,6673 του πλήρους μοντέλου). Τα 9 σταθερά γονίδια – κλειδιά και η κλινική μεταβλητή "Surgery" που επιλέχθηκαν τελικά, καθώς και η βιολογική τους αντιστοίχιση με βάση την πλατφόρμα Agilent, παρουσιάζονται στον Πίνακα 4.4.

Final ID Genes	Final Genes
X54	ABCC11
X2737	CCL13
X4055	CXCL1
X8086	IGSF6
X11667	NRGN
X12376	PCDHB10
X12509	PDIA2
X16872	SUCNR1
X19008	VWA3A

Πίνακας 4. 4 Τελική γονιδιακή υπογραφή των 9 βιοδεικτών που επιλέχθηκαν από το ORSF

4.5 Ποινικοποιημένα Μοντέλα Cox

Για την αντικειμενική αξιολόγηση της απόδοσης των Πλάγιων Τυχαίων Δασών Επιβίωσης, πραγματοποιήθηκε συγκριτική ανάλυση με δύο καθιερωμένες και ιδιαίτερα ισχυρές μεθόδους της βιβλιογραφίας: το ποινικοποιημένο μοντέλο Cox Lasso (Tibshirani (1996)), την αμερόληπτη διαδικασία που προτάθηκε για το ποινικοποιημένο μοντέλο Cox Lasso (Xia et al. (2023)) καθώς και το τροποποιημένο Extreme Gradient Boosting Cox Model (Li et al. (2022)).

4.5.1 Ποινικοποιημένο Μοντέλο Cox Lasso

Το γραμμικό μοντέλο Cox Lasso εκπαιδεύτηκε χρησιμοποιώντας 5 – fold CV για τον προσδιορισμό της βέλτιστης παραμέτρου ποινικοποίησης λ. Με βάση τη βέλτιστη τιμή $\lambda_{min} = 0{,}03855$, ο αλγόριθμος μηδένισε τους συντελεστές 23 γονιδίων, διατηρώντας στο τελικό μοντέλο 23 ενεργά γονίδια με μη – μηδενικούς συντελεστές. Η χρήση της πιο αυστηρής παραμέτρου $\lambda_{1se} = 0{,}17082$ οδήγησε σε υπερβολική απλοποίηση, μηδενίζοντας όλους τους συντελεστές του μοντέλου. Οι εκτιμώμενοι

συντελεστές παλινδρόμησης για τα γονίδια του Cox Lasso παρουσιάζονται στον Πίνακα 4.5.

4.5.2 Ποινικοποιημένο Μοντέλο Cox Debiased Lasso

Η συγκεκριμένη τεχνική ποινικοποίησης εφαρμόσθηκε στην εργασία των Xia et al. (2023) για την αντιμετώπιση της μεροληψίας (bias) που υπάρχει στις εκτιμήσεις των συντελεστών παλινδρόμησης του ποινικοποιημένου μοντέλου Cox Lasso και τη δυνατότητα ελέγχων υποθέσεων για τον εκάστοτε συντελεστή. Το γραμμικό μοντέλο Cox Debiased Lasso εκπαιδεύτηκε με 5 – fold CV για τον προσδιορισμό της βέλτιστης παραμέτρου ποινικοποίησης. Σε επίπεδο σημαντικότητας 5%, προέκυψε ότι 4 μεταβλητές από τις 52 φαίνεται να επιδρούν στο χρόνο επιβίωσης T. Στον Πίνακα 4.5 παρουσιάζονται συγκριτικά τα αποτελέσματα των δύο μοντέλων Lasso (Lasso & Debiased Lasso) για τις 52 μεταβλητές.

ID Gene	coefficient (Cox Lasso)	coefficient (Debiased Cox Lasso)	p-value (Debiased Cox Lasso)	SE (Debiased Cox Lasso)	Active Genes - Debiased Cox Lasso (a=5%)	Active Genes - Cox Lasso
tumor_grade3b	.	-0,49136	0,64006	1,05079	Non-Active	Non-Active
tumor_grade3c	.	-0,13881	0,86310	0,80505	Non-Active	Non-Active
tumor_grade4	.	0,34016	0,67791	0,81905	Non-Active	Non-Active
gradelow	.	-0,05961	0,76043	0,19549	Non-Active	Non-Active
surgerysuboptimal	0,31092	0,88261	0,00009	0,22553	Active	Active
X3251	-0,04977	-0,15920	0,18791	0,12090	Non-Active	Active
X16872	-0,10537	-0,22007	0,15725	0,15560	Non-Active	Active
X2737	-0,05996	-0,06803	0,66597	0,15760	Non-Active	Active
X15025	-0,04066	-0,12392	0,28131	0,11502	Non-Active	Active
X2890	.	0,14255	0,39304	0,16690	Non-Active	Non-Active
X4016	.	-0,03171	0,84017	0,15722	Non-Active	Non-Active
X7273	.	-0,07218	0,61012	0,14155	Non-Active	Non-Active
X15373	.	0,06241	0,69987	0,16189	Non-Active	Non-Active
X2906	.	-0,00282	0,98645	0,16614	Non-Active	Non-Active
X9940	.	0,07093	0,60107	0,13566	Non-Active	Non-Active
X14380	-0,01231	-0,08846	0,48412	0,12643	Non-Active	Active
X12376	0,10321	0,18759	0,06907	0,10319	Non-Active	Active
X2756	.	0,23090	0,57027	0,40676	Non-Active	Non-Active
X6744	.	0,09247	0,49543	0,13565	Non-Active	Non-Active
X11423	.	-0,08309	0,78556	0,30538	Non-Active	Non-Active
X4067	.	-0,00663	0,97902	0,25209	Non-Active	Non-Active
X7494	.	-0,35281	0,36108	0,38630	Non-Active	Non-Active
X12509	0,10786	0,23933	0,04397	0,11881	Active	Active
X13408	.	0,21279	0,40269	0,25428	Non-Active	Non-Active
X1402	-0,01662	-0,10493	0,33314	0,10842	Non-Active	Active
X4849	.	0,00312	0,98967	0,24124	Non-Active	Non-Active
X9465	0,07973	0,09920	0,34613	0,10529	Non-Active	Active
X9812	-0,11117	-0,17525	0,10740	0,10885	Non-Active	Active
X15622	.	-0,23872	0,54778	0,39714	Non-Active	Non-Active

ID Gene	coefficient (Cox Lasso)	coefficient (Debiased Cox Lasso)	p-value (Debiased Cox Lasso)	SE (Debiased Cox Lasso)	Active Genes - Debiased Cox Lasso (a=5%)	Active Genes - Cox Lasso
X2826	.	-0,05859	0,58675	0,10778	Non-Active	Non-Active
X16050	.	-0,29698	0,18117	0,22210	Non-Active	Non-Active
X2125	0,02189	0,05645	0,78802	0,20995	Non-Active	Active
X4389	0,01811	0,06496	0,55145	0,10907	Non-Active	Active
X54	-0,14594	-0,26703	0,01193	0,10621	Active	Active
X7493	.	0,82941	0,14838	0,57387	Non-Active	Non-Active
X9290	-0,17222	-0,22624	0,01319	0,09127	Active	Active
X17473	.	0,00477	0,98513	0,25620	Non-Active	Non-Active
X4055	-0,10307	-0,21425	0,06967	0,11811	Non-Active	Active
X11667	-0,09758	-0,20326	0,07748	0,11513	Non-Active	Active
X18804	.	0,00483	0,97993	0,19214	Non-Active	Non-Active
X266	.	0,01764	0,86340	0,10252	Non-Active	Non-Active
X2988	0,04262	0,14631	0,15654	0,10327	Non-Active	Active
X7504	.	0,35885	0,15235	0,25072	Non-Active	Non-Active
X8086	-0,06916	-0,17985	0,37087	0,20098	Non-Active	Active
X8786	0,08089	0,14612	0,12689	0,09572	Non-Active	Active
X10547	.	-0,05625	0,59554	0,10596	Non-Active	Non-Active
X17191	.	-0,16539	0,68793	0,41177	Non-Active	Non-Active
X19008	-0,09927	-0,13581	0,42347	0,16967	Non-Active	Active
X7496	.	-0,33369	0,32408	0,33839	Non-Active	Non-Active
X9415	.	0,12513	0,52413	0,19644	Non-Active	Non-Active
X12345	0,11617	0,18699	0,05543	0,09762	Non-Active	Active
X12745	-0,08533	-0,20397	0,24156	0,17417	Non-Active	Active

Πίνακας 4. 5 Αποτελέσματα Lasso μοντέλων

Το μοντέλο Cox Lasso επιλέγει συνολικά 23 μεταβλητές σε αντίθεση με το μοντέλο Debiased Lasso που επιλέγει μόλις 4 μεταβλητές σε επίπεδο σημαντικότητας 5%. Παρατηρείται επίσης ότι οι επιλεγμένες μεταβλητές του Debiased Lasso αποτελούν αυστηρό υποσύνολο των 23 μεταβλητών που επέλεξε το Cox Lasso, μειώνοντας περισσότερο από 90% το πλήθος των ενεργών μεταβλητών.

4.6 Εφαρμογή Τεχνικής ML : Μοντέλο XGBoost

Ο αλγόριθμος Extreme Gradient Boosting Cox (XGBoost Cox) εφαρμόστηκε μέσω του πακέτου "Xsurv" που προτάθηκε στην εργασία των Li et al. (2022), πραγματοποιώντας Newton Boosting για τη βελτιστοποίηση της μερικής πιθανοφάνειας Cox. Το μοντέλο εκπαιδεύτηκε με 5–fold CV, 1000 nrounds (εσωτερικές επαναλήψεις) και early stopping (ο αριθμός των επαναλήψεων που ο αλγόριθμος θα σταματήσει την παραγωγή δέντρων εφόσον δεν προκύπτει κάποια βελτίωση στην απόδοση του μοντέλου) ίσο με 5. Για τη διασφάλιση της απόλυτης αναπαραγωγιμότητας των αποτελεσμάτων, ορίστηκε αυστηρά η εκτέλεση του αλγορίθμου σε έναν μοναδικό επεξεργαστικό πυρήνα (single – core execution). Αυτό κρίνεται απαραίτητο καθώς ο Gradient Boosting αλγόριθμος εκτελεί παράλληλους

υπολογισμούς αθροισμάτων κλίσεων (gradients) οι οποίοι μπορούν να οδηγήσουν σε οριακά διαφορετικές τιμές κέρδους (gain) και συνεπώς σε διαφορετική επιλογή γονιδίων σε κάθε εκτέλεση. Με τη διαδικασίας επικύρωσης CV, προέκυψαν οι βέλτιστες υπερπαράμετροι του μοντέλου $\lambda = 0{,}6955,\ \alpha = 0{,}5058\ \&\ \eta = 0{,}0512.$ Έχοντας αυτές τις τιμές, τα τελικά 15 γονίδια στα οποία κατέληξε το μοντέλο XGBoost παρατίθενται στον Πίνακα 4.6.

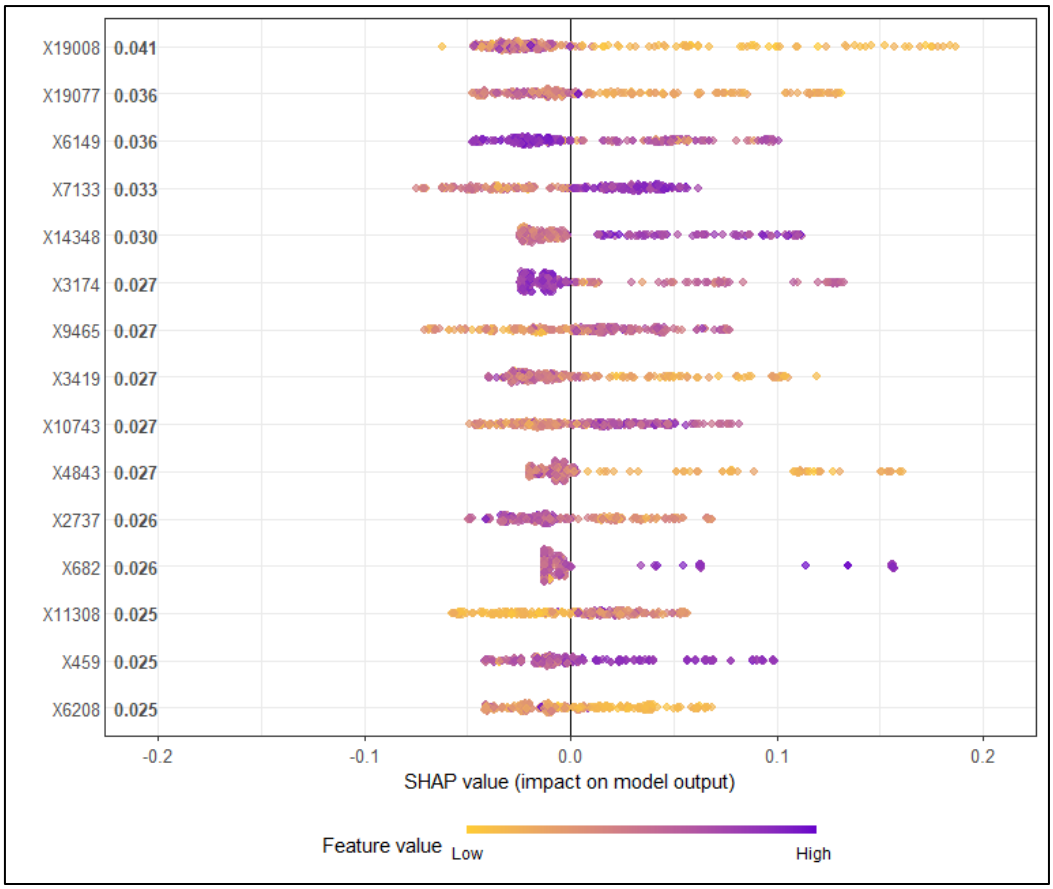
Final ID Genes	Final Genes
X19008	VWA3A
X19077	WDR38
X6149	FOLR3
X7133	GUCY1B2
X14348	RHCE
X3174	CFB
X9465	LINC00669
X3419	CLEC12A
X10743	MRVI1
X4843	DUSP8///DUSP8P3///DUSP8P4///DUSP8P2///DUSP8P1
X2737	CCL13
X682	ANKDD1A
X11308	NELL1
X459	AHNAK2
X6208	FPR2

Πίνακας 4. 6 Η τελική γονιδιακή υπογραφή των 15 βιοδεικτών που επιλέχθηκαν από το XGBoost

Το μοντέλο XGBoost δίνει τη δυνατότητα απεικόνισης της σημαντικότητας των μεταβλητών που επέλεξε. Το γράφημα SHAP (SHapley Additive exPlanations) που παρουσιάζεται στην Εικόνα 4. 5 απεικονίζει την επίδραση των 15 επιλεγμένων γονιδίων στην τελική πρόβλεψη κινδύνου του μοντέλου XGBoost για κάθε ασθενή.

Τα διαθέσιμα στοιχεία του γραφήματος SHAP (Εικόνα 4. 5) είναι:

1) Κατάταξη Σημαντικότητας (Κάθετος Άξονας): Τα γονίδια είναι ταξινομημένα από πάνω προς τα κάτω με βάση τη μέση απόλυτη τιμή SHAP, η οποία αποτυπώνει τη συνολική προγνωστική τους ισχύ. Το γονίδιο X19008 (μέση επίδραση 0,041) αναδεικνύεται ως ο πιο επιδραστικός βιοδείκτης, ακολουθούμενο από το X19077 (0,036) και το X6149 (0,036).
2) Κατεύθυνση της Επίδρασης (Οριζόντιος Άξονας SHAP Value): Οι τιμές δεξιά από το μηδέν ($SHAP > 0$) υποδηλώνουν αύξηση του κινδύνου (χειρότερη επιβίωση – παράγοντας κινδύνου), ενώ οι τιμές αριστερά από το μηδέν ($SHAP < 0$) υποδηλώνουν μείωση του κινδύνου (καλύτερη επιβίωση – προστατευτικός παράγοντας)
3) Επίπεδο έκφρασης (Χρωματική Διαβάθμιση): Το μοβ χρώμα αντιπροσωπεύει υψηλό επίπεδο έκφρασης, ενώ το κίτρινο αντιπροσωπεύει χαμηλό επίπεδο έκφρασης.

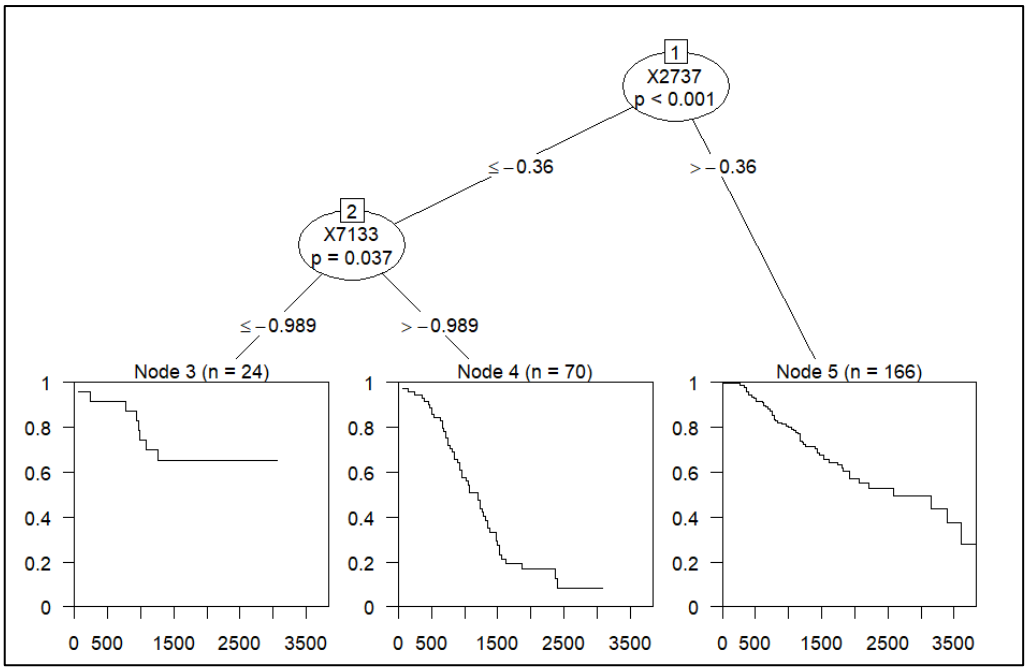


Εικόνα 4. 5 Γράφημα SHAP του μοντέλου XGBoost

Για το γονίδιο X2737, παρατηρείται ότι οι σκουρόχρωμες κουκίδες (υψηλή έκφραση) συγκεντρώνονται στην αριστερή πλευρά ($SHAP < 0$). Αυτό σημαίνει ότι το γονίδιο X2737 δρα ως προστατευτικός παράγοντας στο χρόνο επιβίωσης, μειώνοντας σημαντικά τον κίνδυνο θανάτου του ασθενούς. Αντίθετα, η χαμηλή του έκφραση (ανοιχτόχρωμες κουκίδες) αυξάνει δραματικά τον κίνδυνο.

Σε αντίθεση με το γονίδιο X7133, παρατηρείται ότι οι μοβ κουκίδες (υψηλή έκφραση) συγκεντρώνονται στη δεξιά πλευρά ($SHAP > 0$). Αυτό σημαίνει ότι το γονίδιο X7133 αυξάνει σημαντικά τον κίνδυνο επιβίωσης, ενώ η χαμηλή του έκφραση μειώνει δραματικά τον κίνδυνο.

Ένα από τα κύρια πλεονεκτήματα του πακέτου "Xsurv" είναι η δυνατότητα δημιουργίας ενός απλουστευμένου δέντρου ερμηνείας (Εικόνα 4. 6), αλλά με την κατάλληλη πληροφορία από τον αλγόριθμο XGBoost, έτσι ώστε ο εκάστοτε γιατρός, που το χρησιμοποιεί, να μπορεί να κατατάξει τα άτομα καταλλήλως στις ομάδες που δημιουργούνται. Παρατηρείται ότι το συγκεκριμένο δέντρο καταλήγει στη χρήση μόνο δύο γονιδίων (X2737 & X7133).



Εικόνα 4. 6 Δέντρο Ερμηνείας του XGBoost Xsurv

4.7 Αξιολόγηση Μοντέλων – Συμπεράσματα

Για την αξιολόγηση των παραπάνω μοντέλων, που εκπαιδεύτηκαν βάσει του συνόλου δεδομένων GSE32062, θα γίνει χρήση του συνόλου δεδομένων GSE32063, το οποίο οι Yoshihara et al. (2012) διαχώρισαν από την αρχική έρευνα για τη χρήση του ως σύνολο επικύρωσης των μεθόδων τους. Το συγκεκριμένο σύνολο δεδομένων αποτελείται από τις ίδιες ακριβώς μεταβλητές (κλινικές και γονιδιακές) με πλήθος ασθενών 40 γυναίκες ιαπωνικής καταγωγής. Τα δεδομένα τυποποιήθηκαν για την ορθή υλοποίηση πρόβλεψης του λογαρίθμου του σχετικού κινδύνου (log – relative hazard – risk scores) στα εκπαιδευμένα μοντέλα. Το ποσοστό δεξιάς λογοκρισίας που παρατηρείται στο συγκεκριμένο σύνολο δεδομένων ανέρχεται στο 45% (22 πλήρεις χρόνοι επιβίωσης). Αναφορικά με τις 3 κλινικές μεταβλητές, παρατίθεται ο πίνακας συχνοτήτων τους.

tumor_grade	**grade**	**surgery**
3a: 0	high:17	optimal :19
3b: 3	low :23	suboptimal:21
3c:28		
4 : 9		

Πίνακας 4. 7 Πίνακας Συχνοτήτων των κλινικών μεταβλητών του GSE32063

Τα αποτελέσματα του δείκτη C – Index για κάθε μοντέλο δίνονται στον παρακάτω συγκεντρωτικό πίνακα.

Μέθοδος - Pipeline			Test C - Index
Feature Screening RMST - Stability Selection (25 replications)	Top 47 genes & 3 clinical variables (5 dummy variables)	Cox Lasso Model	0,596
		Debiased Cox Lasso Model	0,614
		ORSF	0,626
XGBoost Algorithm in 3000 genes + 3 clinical variables (5 dummy variables)			0,654

Πίνακας 4. 8 Πίνακας Προβλεπτικής Ικανότητας Μεθόδων (C - Index) στο test dataset GSE32063

Αναφορικά με τις τιμές του δείκτη C – Index που επιτεύχθηκαν στο ανεξάρτητο σύνολο επικύρωσης GSE32063 (κυμαινόμενες από 0,59 έως 0,66), είναι κρίσιμο να επισημανθεί ότι, παρόλο που σε απόλυτους όρους εμφανίζονται μετριοπαθείς, κρίνονται ως απολύτως ικανοποιητικές και πλήρως εναρμονισμένες με τη διεθνή βιβλιογραφία της κλινικής γονιδιωματικής. Η πρόβλεψη της επιβίωσης στον υψηλού βαθμού ορώδη καρκίνο των ωοθηκών αποτελεί ένα εξαιρετικά δυσχερές υπολογιστικό πρόβλημα λόγω της ακραίας βιολογικής και μοριακής ετερογένειας της συγκεκριμένης νεοπλασίας. Επιπλέον, το υψηλό ποσοστό δεξιάς λογοκρισίας που χαρακτηρίζει τα συγκεκριμένα δεδομένα (53,46% στο GSE32062 και 45% στο GSE32063) περιορίζει μαθηματικά τη μέγιστη θεωρητική τιμή που μπορεί να λάβει ο δείκτης συμφωνίας του Harrell. Το γεγονός ότι η προτεινόμενη μεθοδολογία RMST + ORSF επιτυγχάνει C – Index 0,626 χρησιμοποιώντας μια εξαιρετικά ολιγομελή υπογραφή μόλις 10 μεταβλητών, καθώς και η επίδοση του XGBoost (C – Index = 0,654), επιβεβαιώνουν ότι τα μοντέλα έχουν αποκωδικοποιήσει επιτυχώς ένα ισχυρό και κλινικά αξιοποιήσιμο προγνωστικό σήμα, ανάλογο ή και ανώτερο από αντίστοιχες state–of–the–art εφαρμογές μηχανικής μάθησης σε σύνθετα γονιδιωματικά δεδομένα.

Η αξιοπιστία και η προγνωστική ισχύς των αναπτυχθέντων μοντέλων της παρούσας διατριβής επιβεβαιώνεται περαιτέρω μέσω της σύγκρισής τους με τη διεθνή βιβλιογραφία. Συγκεκριμένα, οι Wang & Li (2019), στην εργασία τους για την ανάπτυξη του αλγορίθμου Extreme Learning Machine Cox (ELMCoxBAR), πραγματοποίησαν εκτενή ανάλυση στο ίδιο σύνολο δεδομένων εκπαίδευσης (GSE32062), εφαρμόζοντας state – of – the – art μεθοδολογίες όπως RSF, Cox Lasso και Cox Boost. Στη μελέτη τους, ο διάμεσος δείκτης C – Index κυμάνθηκε σταθερά σε μετριοπαθή επίπεδα μεταξύ 0,58 και 0,62 για όλα τα μοντέλα, αναδεικνύοντας τη δυσκολία πρόβλεψης σε αυτόν τον εξαιρετικά ετερογενή ιστολογικό τύπο. Στη παρούσα διατριβή, η εφαρμογή του pipeline RMSTSIS + ORSF πέτυχε C – Index 0,626, ενώ ο αλγόριθμος XGBoost έφτασε το 0,654. Οι επιδόσεις αυτές κρίνονται ως εξαιρετικά ανταγωνιστικές, καθώς όχι μόνο αγγίζουν ή ξεπερνούν τα benchmarks των Wang & Li (2019), αλλά επιτεύχθηκαν σε ένα πλήρως ανεξάρτητο, εξωτερικό σύνολο επικύρωσης (GSE32063). Η ικανότητα των μοντέλων της παρούσας

διατριβής να διατηρούν υψηλή προβλεπτική ακρίβεια σε εξωτερικά δεδομένα, παρά τη μικρή διάσταση της τελικής γονιδιακής υπογραφής (μόλις 10 μεταβλητές για το ORSF), υπογραμμίζει τη σταθερότητα και τη γενικευσιμότητα της προτεινόμενης μεθοδολογικής προσέγγισης.

4.7.1 Συμπεράσματα

Στόχος του συγκεκριμένου κεφαλαίου είναι η χρήση σύγχρονων τεχνικών για την Ανάλυση Επιβίωσης δεδομένων υψηλής διάστασης μειώνοντας τη διάστασή τους και καταλήγοντας σε ένα ικανοποιητικό μοντέλο πρόβλεψης. Γι' αυτό το σκοπό, εξετάστηκαν συνολικά 4 διαδικασίες – συνδυασμοί τεχνικών. Για τα πρώτα τρία μοντέλα έγινε αρχικά ένα επαναληπτικό φιλτράρισμα μεταβλητών στο σύνολο δεδομένων GSE32062 για την επιλογή των 47 πιο σταθερών γονιδίων, σύμφωνα με τη διαδικασία RMSTSIS που αναλύεται εκτενέστερα στην υποενότητα 2.3.3. Στη συνέχεια, χρησιμοποιώντας μόνο τα 47 γονίδια που προέκυψαν (Πίνακας 4. 2) και τις 3 αρχικές κλινικές μεταβλητές εκπαιδεύτηκαν δύο μοντέλα Cox Lasso, το κλασικό και η αμερόληπτη διαδικασία ποινικοποίησης Lasso που πρότειναν οι Xia et al. (2023), με τα αντίστοιχα αποτελέσματα να παρατίθενται στον Πίνακα 4.5. Από το συγκεκριμένο πίνακα φαίνεται ότι, η Debiased Lasso διαδικασία κατευθύνεται σε ένα πιο αυστηρό μοντέλο πρόβλεψης διατηρώντας μόλις 4 μεταβλητές στατιστικά σημαντικές σε σχέση με το κλασικό μοντέλο Cox Lasso. Αυτό φαίνεται να λειτουργεί υπέρ του Debiased Lasso μοντέλου το οποίο έχει 61,4% προβλεπτική ικανότητα, διαφορά 5,4% ποσοστιαίες μονάδες σε σύγκριση με το κλασικά Cox Lasso.

Στη συνέχεια ακολούθησε η εκπαίδευση ενός βέλτιστου μοντέλου ORSF (του οποίου οι τιμές των υπερπαραμέτρων προέκυψαν μετά από 5 – fold CV διαδικασία) το οποίο οδήγησε σε σχετικά καλύτερο αποτέλεσμα σε σχέση με τα προηγούμενα δύο μοντέλα (C – index = 0,626) ενώ εκτελώντας μία διαδικασία backward elimination κατέληξε στις 10 κορυφαίες μεταβλητές (Πίνακας 4. 4) που φαίνεται να επηρεάζουν το χρόνο επιβίωσης των ατόμων με το συγκεκριμένο τύπο καρκίνου. Οι 10 κορυφαίες μεταβλητές του ORSF φαίνεται να συμφωνούν κατά 75% σε σχέση με τις 4 που επέλεξε το Debiased Lasso μοντέλο ("Surgery", "X54" & "X12509") ενώ τα μοντέλα ORSF και το Cox Lasso εμφανίζουν πλήρη συμφωνία στις επιλεγμένες μεταβλητές τους με το Cox Lasso να έχει επιλέξει παραπάνω 13 μεταβλητές. Είναι αξιοσημείωτο το γεγονός ότι υπάρχει αρκετά ικανοποιητική συμφωνία των 3 μεθόδων με την ORSF να εμφανίζει την καλύτερη μεταξύ τους προγνωστική ικανότητα.

Σχετικά με το μοντέλο XGBoost παρατηρείται ότι η τεχνική φιλτραρίσματος RMSTSIS έχει 3 κοινές μεταβλητές σε σχέση με τις κορυφαίες 15 που επέλεξε το XGBoost από τις 3.000 γονιδιακές μεταβλητές που περιείχε (X2737: "CCL13", X9465: "LINC00669" & X19008: "VWA3A") ενώ με το μοντέλο Cox Lasso έχει αυτές πάλι τις 3 μεταβλητές, με το Debiased Lasso καμία κοινή και με το ORSF δύο κοινές μεταβλητές (X2737 & X19008). Αξιοσημείωτη είναι η θετική σημαντικότητα του γονιδίου X2737 το οποίο εμφανίζεται στο Δέντρο Ερμηνείας του XGBoost ως ο πρώτος βασικός διαχωρισμός του και αποτελεί έναν προστατευτικό παράγοντα για

τον καρκίνο των ωοθηκών. Σημαντικό γονίδιο για το χρόνο επιβίωσης φαίνεται να αποτελεί και για το ORSF μοντέλο.

Το XGBoost εμφάνισε μεν τη μέγιστη προβλεπτική ικανότητα στο ανεξάρτητο σύνολο επικύρωσης (C – Index = 0,654), ωστόσο η εγγενής στοχαστική φύση της εκπαίδευσής του (μέσω τυχαίας υποδειγματοληψίας παρατηρήσεων και μεταβλητών σε κάθε boosting round) εισάγει έναν βαθμό μεταβλητότητας στα τελικά αποτελέσματα. Επιπλέον, η απευθείας εισαγωγή ενός τόσο εκτενούς γονιδιακού προφίλ (3.000 γονίδια) χωρίς την προηγούμενη εφαρμογή κάποιου αυστηρού φιλτραρίσματος σταθερότητας (όπως η σταθερή επιλογή RMSTSIS), καθιστά τον αλγόριθμο εξαιρετικά ευάλωτο στον κίνδυνο έντονης υπερπροσαρμογής (overfitting) στα δεδομένα, περιορίζοντας τη σταθερότητα και την κλινική του χρησιμότητα. Αντίθετα, ο συνδυασμός του φιλτραρίσματος RMSTSIS με την επακόλουθη εφαρμογή της τεχνικής ORSF (η οποία εισάγει ποινή Elastic Net σε κάθε κόμβο) αναδεικνύεται ως η βέλτιστη προσέγγιση. Η μεθοδολογία αυτή επιτυγχάνει ένα σταθερό, ολιγομελές και κλινικά ερμηνεύσιμο προγνωστικό μοντέλο με τη δυνατότητα επιλογής μεταβλητών, θωρακίζοντάς το αποτελεσματικά τόσο απέναντι σε προβλήματα πολυσυγγραμμικότητας (μέσω της ποινής Ridge) όσο και απέναντι στον κίνδυνο υπερπροσαρμογής.

Παράρτημα – Πίνακες

Ordered selection of variables \ Iteration	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1	X16872	X3251	X15025	X4016	X3251	X3251	X3251	X16872	X15025	X3251	X1402	X3251	X16872	X3251	X3251	X266	X4067	X16872	X16872	X6744	X3251	X6744	X16872	X8786	X11863
2	X14380	X15025	X14974	X3251	X54	X15025	X6744	X3251	X14380	X1402	X6744	X6744	X4849	X6744	X14380	X4016	X3251	X7494	X2125	X16872	X3436	X16872	X3251	X7273	X3251
3	X16065	X4055	X6177	X7493	X6744	X16872	X2737	X12376	X9812	X16872	X15373	X16872	X6158	X4016	X15025	X3251	X2737	X7493	X12345	X16109	X9765	X2831	X7273	X16050	X9465
4	X7273	X4588	X2737	X16872	X8786	X15373	X16050	X14380	X16872	X9812	X18890	X15373	X9290	X16872	X54	X16872	X15025	X6744	X3251	X1843	X2737	X15373	X18416	X4389	X7504
5	X2656	X2737	X9940	X7494	X16872	X12509	X16872	X266	X7273	X6744	X18892	X19925	X3251	X7493	X4588	X2737	X2906	X9904	X9465	X9904	X2799	X12376	X4067	X54	X15025
6	X19101	X15373	X4016	X7273	X4849	X2737	X12376	X7746	X3251	X2890	X2737	X12345	X12509	X7494	X2906	X14380	X2756	X4016	X2890	X8718	X7273	X3251	X1402	X19966	X2906
7	X4055	X1509	X4055	X7504	X11667	X16050	X9465	X4067	X266	X4016	X2826	X2737	X15025	X13408	X4016	X1270	X7494	X8086	X6158	X16050	X2906	X4016	X17473	X15373	X7493
8	X12376	X7273	X4230	X2906	X2125	X4055	X9812	X15837	X9465	X11863	X3251	X4849	X2125	X2868	X16872	X15512	X15622	X2890	X17013	X17013	X2757	X2756	X9955	X15025	X6059
9	X2890	X8184	X7494	X7496	X525	X6744	X15373	X9465	X7915	X5413	X266	X12376	X15373	X9465	X2737	X4849	X2868	X12376	X19008	X1402	X17191	X11175	X15373	X9415	X3433
10	X10526	X14380	X3571	X5706	X4016	X1402	X14380	X2890	X1402	X15373	X16050	X15025	X12822	X15373	X9940	X11308	X17473	X9812	X7493	X4389	X2756	X2890	X2880	X10547	X9765
11	X19321	X673	X2826	X11423	X7133	X2757	X7915	X15373	X11667	X2906	X12509	X11423	X6744	X2756	X3436	X11423	X2869	X14380	X7273	X11667	X8805	X2868	X4016	X14380	X7494
12	X9940	X2890	X2656	X18804	X16050	X5943	X13408	X10535	X7746	X4204	X10022	X3433	X2826	X1402	X16485	X7494	X2757	X3571	X2737	X10547	X9312	X7273	X2906	X4849	X11667
13	X2019	X2906	X7493	X7169	X11423	X9415	X8718	X9812	X2890	X8114	X9465	X11308	X8786	X11423	X11423	X7493	X16872	X7496	X11903	X8078	X18416	X15025	X2890	X12376	X7273
14	X10943	X266	X15373	X8114	X12376	X14380	X14974	X4204	X3436	X2737	X15025	X2906	X54	X2890	X16050	X7746	X7493	X2737	X4849	X546	X2868	X8086	X15622	X9465	X15656
15	X8815	X12509	X2890	X6467	X18032	X8184	X2756	X9940	X5952	X18416	X5952	X8549	X9906	X16050	X13408	X425	X7959	X16050	X7494	X713	X2869	X17473	X2831	X16872	X2890
16	X3251	X9765	X17246	X4067	X9415	X525	X9415	X2737	X2826	X15025	X2125	X6840	X11903	X15622	X4067	X2125	X17191	X2756	X14492	X3014	X16872	X8718	X2869	X19008	X266
17	X2737	X10547	X14492	X13408	X14380	X2890	X2890	X7273	X7494	X14380	X2988	X11667	X2988	X17473	X9290	X13408	X2872	X2988	X9955	X4849	X11423	X19008	X2756	X2125	X15622
18	X266	X8086	X2019	X9955	X7273	X1476	X6467	X2125	X6744	X7915	X3571	X9465	X2799	X17191	X17473	X2756	X2837	X6908	X54	X6177	X3433	X14380	X13408	X9940	X7959
19	X18946	X17473	X3251	X7959	X2737	X12376	X4588	X4016	X16109	X2826	X9415	X16050	X9955	X12376	X2756	X2906	X2901	X8381	X15409	X9906	X15373	X14364	X132	X1839	X3411
20	X12554	X16009	X9290	X2756	X15512	X13408	X1402	X3753	X2737	X10547	X19008	X8786	X10547	X7496	X12745	X2890	X54	X11175	X13161	X19966	X13408	X5943	X8906	X17246	X4067
21	X2295	X1975	X4588	X9465	X3941	X4849	X2656	X15025	X3433	X17876	X11667	X12509	X7494	X7273	X19101	X11667	X9290	X3829	X9940	X9415	X17473	X2125	X18101	X10927	X9940
22	X2613	X4016	X19077	X2737	X4133	X4389	X2757	X2756	X9940	X4956	X1839	X2890	X19804	X2869	X7915	X8902	X4059	X2906	X15025	X9812	X3060	X15459	X9191	X6744	X9833
23	X8786	X15489	X16009	X12345	X11863	X546	X4016	X12509	X9290	X9940	X3436	X8086	X16366	X15025	X18416	X11192	X8381	X8078	X12386	X14380	X8086	X9940	X8786	X3251	X19008
24	X4230	X15622	X1586	X5413	X2988	X4016	X2438	X2906	X5413	X7273	X16109	X2756	X11	X4067	X6177	X7504	X15062	X8902	X16009	X16994	X18804	X8078	X17191	X15837	X1586
25	X4647	X13408	X5951	X15622	X940	X2756	X2906	X16991	X54	X17473	X4055	X1402	X4389	X2901	X15622	X9812	X7273	X3884	X18804	X15373	X15622	X2881	X15062	X4055	X12345
26	X15373	X7746	X15410	X12509	X2906	X2125	X8786	X18416	X7504	X13408	X4849	X2988	X3773	X7959	X2826	X4067	X4074	X5951	X12376	X18890	X9940	X12509	X8148	X6177	X10547
27	X12745	X16872	X19099	X15512	X4067	X2988	X2826	X11423	X4588	X12509	X7273	X4016	X4204	X6467	X2438	X2988	X18111	X14364	X15837	X54	X7168	X18804	X15025	X8718	X7496
28	X14315	X2826	X19008	X11175	X7915	X520	X17191	X7959	X4055	X12376	X1407	X4389	X4016	X3436	X18032	X15622	X11423	X5943	X2988	X10927	X8148	X13408	X2737	X3411	X4588
29	X6177	X3753	X16872	X3411	X9955	X12386	X12745	X18804	X15373	X9904	X5951	X9940	X13742	X7896	X12509	X12345	X9116	X15512	X19966	X18892	X11667	X4067	X54	X16009	X4063
30	X17905	X15459	X9812	X2831	X5312	X17862	X4849	X4056	X7493	X16109	X2890	X18562	X1312	X7504	X1407	X12376	X15410	X4067	X7504	X2969	X2872	X4849	X16295	X11308	X9812
31	X17449	X19077	X12745	X2890	X2756	X2745	X7273	X2834	X19077	X4389	X1899	X2656	X7273	X2737	X19966	X6751	X6467	X4729	X15533	X3251	X9116	X1402	X15631	X9812	X17473
32	X8184	X9290	X8086	X6908	X7020	X10535	X12646	X9906	X8908	X6371	X3558	X2826	X9904	X18416	X9415	X1402	X16009	X8905	X7902	X9908	X3245	X8786	X12509	X2656	X2799
33	X18376	X2745	X11863	X7168	X12745	X4067	X8549	X5943	X10547	X9290	X17625	X2745	X11423	X7168	X11175	X132	X4016	X16109	X14315	X17349	X4069	X7494	X9465	X12509	X18557
34	X12345	X9812	X19321	X9833	X13408	X8815	X6158	X3434	X8563	X9443	X9908	X3565	X7680	X132	X2019	X8903	X2793	X132	X7496	X9940	X14364	X4389	X58	X1402	X9290
35	X9290	X12646	X14380	X6744	X2826	X9803	X9765	X15622	X12551	X2880	X4389	X5943	X3941	X2831	X19008	X7496	X7896	X14315	X9906	X2125	X7896	X15512	X7493	X6059	X4055
36	X2745	X3433	X17905	X8086	X4389	X9940	X15622	X6466	X16984	X2901	X9812	X5951	X1839	X8718	X2869	X18804	X11863	X18804	X2838	X7494	X9290	X6840	X9765	X8078	X18944
37	X12386	X1586	X2744	X6751	X6751	X4056	X11423	X2833	X4067	X10744	X546	X7496	X266	X16991	X4849	X17191	X8086	X7504	X525	X15025	X2758	X2906	X11423	X4204	X4016
38	X4700	X8161	X4389	X14315	X7494	X16991	X18890	X2831	X19966	X18032	X2500	X2880	X4055	X8381	X3411	X9465	X1839	X11308	X13742	X2737	X8846	X2737	X5951	X8908	X2868
39	X429	X7504	X4764	X9940	X11175	X18804	X2744	X17191	X2906	X11423	X12376	X9765	X16991	X2837	X2656	X54	X2826	X9940	X13052	X6158	X6177	X18890	X632	X12745	X17625

Πίνακας 4. 9 Αποτελέσματα επαναληπτικής διαδικασίας σταθερής επιλογής μεταβλητών με RMSTSIS

Variable Selection by ORSF (Permutation Importance)		
N Predictors	**C-Index**	**Dropped Predictor**
10	0,69727	NA
11	0,70912	X12345
12	0,70719	X9812
13	0,71421	X7273
14	0,71273	X12745
15	0,70917	X9290
16	0,71055	X4849
17	0,70439	X15025
18	0,70058	X2125
19	0,69504	X14380
20	0,70149	X17191
21	0,69824	X10547
22	0,69438	X1402
23	0,70373	X8786
24	0,70399	X11423
25	0,69956	X266
26	0,69966	X3251
27	0,69590	X9465
28	0,69356	X9940
29	0,68634	X15373
30	0,69443	X4067
31	0,69300	X2988
32	0,69163	X15622
33	0,68797	X2906
34	0,69061	X2756
35	0,69178	tumor_grade_3b
36	0,68411	X4389
37	0,67882	X17473
38	0,67617	X18804
39	0,67877	X2826
40	0,68177	X16050
41	0,67190	tumor_grade_4
42	0,67012	X6744
43	0,66885	X7496
44	0,67241	X9415
45	0,67190	tumor_grade_3c
46	0,66646	X4016
47	0,66392	X7494
48	0,66418	X2890
49	0,66814	grade_low
50	0,66590	X13408
51	0,66011	X7493
52	0,66728	X7504

Πίνακας 4. 10 Διαδικασία Variable Selection για την εύρεση των 10 κορυφαίων μεταβλητών στο τελικό μοντέλο ORSF

Παράρτημα – Κώδικας Εργασίας στη γλώσσα προγραμματισμού R

```r
## ============================================================================
## High Dimensional Survival Analysis: Methods' Application in Training Dataset
## ============================================================================

# ============================================================================
#                            Preprocess Training Data
# ============================================================================
# Libraries
library(BiocManager)
library(curatedOvarianData)

# Load and save dataset GSE32062
data("GSE32062.GPL6480_eset", package = "curatedOvarianData")
eset <- GSE32062.GPL6480_eset

# Pre-processing genomic variables
genomic_data <- data.frame(t(exprs(eset)),check.names = F)

# Pre-processing time and status about GSE32062
pheno <- pData(eset)
time <- pheno$days_to_death
status <- ifelse(pheno$vital_status=="living",0,1)

# Pre-processing clinical variables
pheno$substage <- ifelse(is.na(pheno$substage),"",pheno$substage)
clinical_var <- data.frame("tumor_grade"= as.factor(paste(pheno$tumorstage,pheno$substage,sep="")),
                           "grade" = as.factor(pheno$summarygrade),
                           "surgery" = as.factor(pheno$debulking))

# Rename the names of genomic variables
GSE32062 <- data.frame(time,status,clinical_var,genomic_data)
genes_name <- colnames(genomic_data)
colnames(GSE32062)[-c(1:5)] <- paste("X",1:ncol(genomic_data),sep = "")

# ============================================================================
#                              Descriptive Analysis
# ============================================================================
# Descriptive Measures
library(psych)

summary_genomic <- describe(GSE32062[,-c(1:5)], quant = c(.25,.75), IQR = T, type = 2)
hist(GSE32062$time, main = "Histogram of Overall Survival Time", xlab = "Time (in days)", 20)
summary(GSE32062[,c(2:5)])

# Kaplan - Meier Curves
library(survival)
library(survminer)

# Overall Time and Status
fit <- survfit(Surv(time, status) ~ 1, data = GSE32062)
ggsurvplot(fit, data = GSE32062, legend = "none",
           title = "Kaplan-Meier Curve for Overall Survival",
           xlab = "Survival Time (in days)",
           ylab = "Survival Probability",
           conf.int = T, risk.table = T,palette = "navyblue",surv.median.line = "hv")

# Time, Status in Tumor Stage
fit1 <- survfit(Surv(time, status) ~ tumor_grade, data = GSE32062)
p1 <- ggsurvplot(fit1, data = GSE32062,
           title = "Kaplan-Meier Curve in Tumor Grade",
           xlab = "Survival Time (in days)",
           ylab = "Survival Probability")

p1$plot <- p1$plot +
```

```r
  guides(color = guide_legend(nrow = 2, byrow = TRUE, title = NULL))

# Time, Status in Grade
fit2 <- survfit(Surv(time, status) ~ grade, data = GSE32062)
p2 <- ggsurvplot(fit2, data = GSE32062,
           title = "Kaplan-Meier Curve in Grade",
           xlab = "Survival Time (in days)",
           ylab = "Survival Probability")

p2$plot <- p2$plot +
  guides(color = guide_legend(nrow = 1, byrow = TRUE, title = NULL))

# Time, Status in Surgery
fit3 <- survfit(Surv(time, status) ~ surgery, data = GSE32062)
p3 <- ggsurvplot(fit3, data = GSE32062,
           title = "Kaplan-Meier Curve in Surgery",
           xlab = "Survival Time (in days)",
           ylab = "Survival Probability")

p3$plot <- p3$plot +
  guides(color = guide_legend(nrow = 1, byrow = TRUE, title = NULL))


# Arrange them in 1 row and 3 columns (side-by-side)
library(gridExtra)
grid.arrange(p1$plot,p2$plot,p3$plot, ncol = 3)

# ============================================================================
#                     Cleaning Data by Median Absolute Deviation
# ============================================================================

mad_values <- as.matrix(sort(apply(GSE32062[,-c(1:5)],2,mad),decreasing = T))

# ============================================================================
#                    Final Genomic Dataset for analysis
# ============================================================================

x <- GSE32062[,rownames(mad_values)[1:3000]]

data <- data.frame(GSE32062[,c(1:5)],scale(x)) #final data for analysis

# ============================================================================
# ============================================================================
# ============================================================================

# ============================================================================
#    Feature Screening in Genomic Variables by RMST - Stability Selection
# ============================================================================

# Setup reproducibility
set.seed(42)

# Basic parameters setup
n <- nrow(data)    # Total number of observations (patients)
p <- ncol(x)       # Total number of covariates (genes/features)

# Load the RMST screening functions
# The function "RMSTSIS" is uploaded in "Supplementary Material for 'High-dimensional
# Feature Screening for Nonlinear Associations With Survival Outcome Using Restricted
# Mean Survival Time'" by Yaxian Chen, K.F. Lam, and Zhonghua Liu
# DOI: https://doi.org/10.1002/sta4.673

source("RMST.R")
```

```r
# Define training set size (80% of total samples)
n_train <- ceiling(0.8*nrow(data))

# Screening threshold: d = ceiling(n_train / log(n_train))
p_train <- ceiling(n_train / log(n_train))  # For n_train = 208, this is 39 genes

# Initialize a matrix to record selected gene indices for stability selection (p_train rows, 25 columns)
n_reps <- 25
selected_genes <- matrix(0, nrow = p_train, ncol = n_reps)

# Start the 25-replication loop
for(k in 1:n_reps){
  # Randomly split the data (80% Train / 20% Test)
  train_id <- sample(1:n, size = n_train)
  x_train  <- data[train_id, -c(1:5)]
  y_train  <- data[train_id, c("time", "status")]

  # Run RMST screening on the training set
  # RMSTSIS returns indices of selected covariates sorted by significance
  selected_idx <- RMSTSIS(x = x_train, y = y_train$time, del = y_train$status, thres = p_train)

  selected_genes[, k] <- selected_idx

  cat("The iteration ",k," finished!\n")
}

# Define the names of variables of "selected genes" with real names because now algorithm return the index in dataset
gene_names_pool <- colnames(data)[-c(1:5)]

selected_genes_names <- matrix(gene_names_pool[selected_genes],
                               nrow = nrow(selected_genes),
                               ncol = ncol(selected_genes))

# Top 47 variables which are selected based on frequency via the aforementioned replication process
final_thres <- ceiling(nrow(data)/log(nrow(data)))

freq_genes <- sort(table(as.vector(selected_genes_names)),decreasing = T)

top47 <- names(freq_genes)[1:final_thres]

# ============================================================================
#                  Correlation Spearman between top 47 genes
# ============================================================================

correl <- cor(data[,top47],method = "spearman")

library(ggplot2)
library(ggcorrplot)

ggcorrplot(correl,method="square",type="upper")+
  theme(axis.text.x = element_text(size=9,angle = 90),
        axis.text.y = element_text(size=9))

# ============================================================================
#                           Cox Lasso Model
# ============================================================================
library(glmnet)

# Define the data suitably for glmnet
y_lasso <- Surv(data$time,data$status)

clinical_vars <- colnames(data)[3:5]
gene_vars <- top47

formula_str <- paste("~", paste(c(clinical_vars, gene_vars), collapse = " + "))
```

```r
x_lasso <- model.matrix(as.formula(formula_str), data = data)[, -1]

n_clinical_cols <- ncol(x_lasso) - length(gene_vars)

# Fit Linear Cox Lasso using cross-validation to find optimal lambda
set.seed(42)

cv_cox_lasso <- cv.glmnet(
  x = x_lasso,
  y = y_lasso,
  family = "cox",
  alpha = 1,
  nfolds = 5,
  cox.ties = "efron",
  standardize = F)

# Retrieve coefficients at the optimal lambda.min
coef_min <- coef(cv_cox_lasso, s = "lambda.min")

# Print only the genes that have non-zero coefficients
active_coefs <- coef_min[as.vector(coef_min) != 0, , drop = FALSE]; active_coefs



# ================================================================================
#                           Debiased Lasso Cox Model
# ================================================================================
# The following code is from paper '2021 Debiased Lasso - Statistical inference
# for Cox proportional hazards models' by Xia Lu, Nan B, Li Y. and is found in
# Github: https://github.com/luxia-bios/CoxInference/
library(quadprog)
library(survival)
library(glmnet)
library(Rcpp)

## Source C++ Helper Functions
sourceCpp("sim_univLib.cpp")

## Data Preprocessing & Variable Alignment
# Clinical adjustment variables (located in columns 3, 4, and 5 of 'data')
clinical_vars <- colnames(data)[3:5]

# Retrieve gene names from your top47 indices
gene_vars_raw <- top47

# Generate the model matrix including clinical dummies (no intercept) and genomic features
# Backticks are added around variable names to handle non-syntactic characters safely
safe_vars <- paste0("`", c(clinical_vars, top47), "`")
formula_str <- paste("~", paste(safe_vars, collapse = " + "))

X <- model.matrix(as.formula(formula_str), data = data)[, -1]
time <- data$time
delta <- data$status

## Re-order the training data in ascending order of survival time
order_idx <- order(time)
X <- X[order_idx, ]
delta <- delta[order_idx]
time <- time[order_idx]

## Setup Hyperparameters
n_multi <- 30 # Number of gamma_n values tried in cross-validation
multiplier_seq <- exp(seq(from = log(0.01), to = log(3), length.out = n_multi)) # Gamma_n sequence
n_lambda <- 100 # Number of lambda values for lasso in glmnet
nfold <- 5      # Cross-validation folds for lasso in glmnet
alpha <- sig_level <- 0.05
```

```r
alpha_cv <- 0.1
v <- qnorm(sig_level/2, lower.tail = FALSE)
tol <- 1.0e-6
maxiter <- 50000

## Cross-Validation for Lasso Estimator
set.seed(42)
cvobj_glmnet <- cv.glmnet(x = X,y = cbind(time = time, status = delta),family = "cox",
  alpha = 1, standardize = FALSE, nfolds = nfold, nlambda = n_lambda, cox.ties = "efron")

# Fit Lasso at optimal lambda.min
beta_glmnet <- as.vector(coef(glmnet(x = X, y = cbind(time = time, status = delta), family = "cox",
                                     alpha = 1, lambda = cvobj_glmnet$lambda.min, standardize = FALSE,
                                     thresh = tol, maxit = maxiter, cox.ties = "efron")))

## Calculate Hessian and Derivatives (Using C++ Functions)
p <- ncol(X)
n <- nrow(X)
neg_loglik_glmnet <- 0                      # Log partial likelihood
neg_dloglik_glmnet <- rep(0, p)              # First-order derivative
neg_ddloglik_glmnet <- matrix(0, nrow=p, ncol=p) # Second-order derivative
score_sq <- matrix(0, nrow=p, ncol=p) # Sigma matrix (unbiased Hessian)

# Call C++ function evaluated at the lasso estimate
neg_loglik_functions_cpp_ext(neg_loglik_glmnet, neg_dloglik_glmnet, neg_ddloglik_glmnet, score_sq, X, time, delta,
beta_glmnet)

# Clean small eigen values due to numerical precision
r <- eigen(score_sq)
r$values[r$values <= 1.0e-14] <- 0

## new method ##

## Cross-Validation for Tuning Parameter Gamma_n (via Quadratic Programming)
n_cv <- 5 # number of cv folds for Theta matrix estimation in QP
all_cv_idx <- rep(1:n_cv, length.out = n)
set.seed(42)
all_cv_idx <- sample(all_cv_idx, size = n)
# Define variable for cross-validation criterion values
all_cvpl2 <- array(NA, length(multiplier_seq)) # lik evaluation on left-out data

for(jj in 1:length(multiplier_seq)) {# for cvpl
  multiplier <- multiplier_seq[jj]
  # cv
  cv_idx <- NULL
  cvpl2 <- 0

  for(k in 1:n_cv) {# for cv
    cv_idx <- which(all_cv_idx == k) # test data
    train_x <- X[-c(cv_idx), ]
    test_x  <- X[cv_idx, ]
    train_time <- time[-c(cv_idx)]
    test_time <- time[cv_idx]
    train_delta <- delta[-c(cv_idx)]
    test_delta  <- delta[cv_idx]
    # use training data to obtain de-biased estimator at a given gamma_n
    set.seed(42)
    cvobj_glmnet_train <- cv.glmnet(x = train_x, y = cbind(time = train_time, status = train_delta), family = "cox",
                                    alpha = 1, standardize = FALSE, nfolds = nfold, nlambda = n_lambda, cox.ties = "efron")
    beta_glmnet_train <- as.vector(coef(glmnet(x = train_x, y = cbind(time = train_time, status = train_delta),
                                               family = "cox", alpha = 1, lambda = cvobj_glmnet_train$lambda.min,
                                               standardize = FALSE, thresh = tol,maxit = maxiter, cox.ties = "efron")))

    # Re-calculate derivatives for the train partition
    neg_loglik_glmnet_train <- 0
    neg_dloglik_glmnet_train <- rep(0, p)
```

```
    neg_ddloglik_glmnet_train <- matrix(0, nrow=p, ncol=p)
    score_sq_train <- matrix(0, nrow=p, ncol=p)

    neg_loglik_functions_cpp_ext(neg_loglik_glmnet_train, neg_dloglik_glmnet_train, neg_ddloglik_glmnet_train,
      score_sq_train, train_x, train_time, train_delta, beta_glmnet_train)

    r_train <- eigen(score_sq_train)
    r_train$values[r_train$values <= 1.0e-14] <- 0

    ### CV: QP (using solve.QP) ###

    b_hat_new <- rep(NA, p)
    se_new <- rep(NA, p)

    # Calculate partition-specific mu (gamma_n)
    mu <- multiplier * sqrt(log(p) / nrow(train_x))

    my_pos <- which(r_train$values > 0)
    my_rank <- sum(r_train$values > 0)

    Dmat <- diag(r_train$values[my_pos])
    dvec <- rep(0, my_rank)
    Amat <- t(rbind(-r_train$vectors[, my_pos] %*% Dmat, r_train$vectors[, my_pos] %*% Dmat))

    for(j in 1:p) {
      e_j <- rep(0, p); e_j[j] <- 1
      bvec <- (c(-e_j, e_j) - mu * rep(1, 2 * p))
      res <- solve.QP(Dmat = Dmat, dvec = dvec, Amat = Amat, bvec = bvec)
      m <- as.vector(r_train$vectors[, my_pos] %*% res$solution) +
        as.vector(r_train$vectors[, -my_pos] %*% rep(0, p - my_rank))
      b_hat_new[j] <- beta_glmnet_train[j] - as.numeric(m %*% neg_dloglik_glmnet_train)
      se_new[j] <- sqrt(m[j] / nrow(train_x))
    }

    # Obtain hard thresholded de-biased lasso estimator for de-noising for CV evaluation
    pval_new <- 2 * pnorm(abs(b_hat_new / se_new), lower.tail = FALSE)
    tmp_beta <- b_hat_new * as.numeric(pval_new < (alpha_cv / p))
    cvpl2 <- cvpl2 + loglik_cpp_ext(X = test_x, time = test_time, delta = test_delta, beta = tmp_beta)
  }# end for cv
  all_cvpl2[jj] <- cvpl2
}# end for cvpl

##### results using the chosen tuning parameter from all_cvpl2 #####
optimal_idx <- which.max(all_cvpl2)
multiplier <- multiplier_seq[optimal_idx] # chosen multiplier

b_hat_new <- array(NA, p) # this is the final estimator b_hat
se_new <- array(NA, p) # this is the model-based SE for b_hat
theta_new <- matrix(NA, ncol = p, nrow = p) # estimated Theta matrix from QP approach
mu_new <- multiplier * sqrt(log(p) / n) # chosen tuning parameter gamma_n
my_pos <- which(r$values > 0)
my_rank <- sum(r$values > 0)
Dmat <- diag(r$values[my_pos])
dvec <- rep(0, my_rank)
Amat <- t(rbind(-r$vectors[, my_pos] %*% Dmat, r$vectors[, my_pos] %*% Dmat))

for(j in 1:p) {
  e_j <- rep(0, p); e_j[j] <- 1
  bvec <- (c(-e_j, e_j) - mu_new * rep(1, 2 * p))
  res <- solve.QP(Dmat = Dmat, dvec = dvec, Amat = Amat, bvec = bvec)
  m <- as.vector(r$vectors[, my_pos] %*% res$solution) +
    as.vector(r$vectors[, -my_pos] %*% rep(0, p - my_rank))
  b_hat_new[j] <- beta_glmnet[j] - as.numeric(m %*% neg_dloglik_glmnet)
  se_new[j] <- sqrt(m[j] / n)
  theta_new[j, ] <- m
```

```r
}

# Final p-values
pval_new  <- 2 * pnorm(abs(b_hat_new / se_new), lower.tail = FALSE)

## Results Debiased Cox Lasso & Cox Lasso
results_lasso <- data.frame("ID Gene" = colnames(X),
                            "coefficient (Cox Lasso)" = round(as.vector(unname(coef_min)),7),
                            "coefficient (Debiased Cox Lasso)" = round(b_hat_new,7),
                            "p-value (Debiased Lasso)" = round(pval_new,7),
                            "SE (Debiased Lasso)" = round(se_new,7),
                            "Active Genes - Debiased Lasso (a=5%)" = ifelse(pval_new<0.05,"Active","Non-Active"),
                           "Active Genes - Cox Lasso" = ifelse(as.vector(unname(coef_min))!=0,"Active","Non-Active"),
                            check.names = F)

# ================================================================================
#   Oblique Random Survival Forest with 47 genes and clinical variables
# ================================================================================

## Grid Search for best alpha and mtry in ORSF with top 47 genes and clinical variables
# Input
library(survival)
library(aorsf)
library(dplyr)

# GSE32062 with top 47 variables and clinical variables
X_32062 = data[,c("tumor_grade","grade","surgery",top47)]
time_32062 = data$time
status_32062 = data$status

# Settings
n_folds <- 5
n_tree <- 1000

# Parameter grids
alpha_grid <- c(0, 0.25, 0.50, 0.75, 1)
mtry_grid <- c(5, 10, 15, 20)

# Create 5 Stratified CV folds
# Create folds separately within event / censored groups
set.seed(42)
fold_id <- integer(nrow(X_32062))

event_idx <- which(status_32062 == 1)
cens_idx  <- which(status_32062 == 0)

fold_id[event_idx] <- sample(rep(1:n_folds, length.out = length(event_idx)))
fold_id[cens_idx] <- sample(rep(1:n_folds, length.out = length(cens_idx)))

# Check number of events in each fold
cat("\nEvents per fold:\n")
print(table(fold_id, status_32062))

# Storage for results
cv_results <- list()
counter <- 1

# 5-fold CV
for (fold in 1:n_folds) {
  # Train / validation indices
  train_idx <- which(fold_id != fold)
  valid_idx <- which(fold_id == fold)

  X_train <- X_32062[train_idx, ,drop = FALSE]
  X_valid <- X_32062[valid_idx, ,drop = FALSE]
```

```r
  time_train <- time_32062[train_idx]
  status_train <- status_32062[train_idx]

  time_valid <- time_32062[valid_idx]
  status_valid <- status_32062[valid_idx]

  # Tuning Grid
  for (mtry_value in mtry_grid) {
    for (alpha_value in alpha_grid) {
      cat("Testing mtry = ",mtry_value,"| alpha = ",alpha_value,"\n")

      # ORSF control
      control_net <- orsf_control_survival(method="net",scale_x = F,ties="efron",net_mix=alpha_value)

      # Training data
      train_data <- data.frame(time = time_train,status = status_train,X_train,check.names = FALSE)
      valid_data <- data.frame(X_valid, check.names = FALSE)

      # ORSF
      set.seed(42)
      fit <- orsf(formula = Surv(time, status) ~ ., data = train_data, mtry = mtry_value, attach_data = F,
                  n_tree = n_tree, control = control_net, verbose_progress = T,importance = "none")

      # Prediction
      risk_pred <- predict(fit, new_data = valid_data)

      # C-index
      cindex <- concordance(Surv(time_valid, status_valid) ~ risk_pred, reverse = TRUE)$concordance
      cat("C - Index = ",cindex,"\n")

      # Save result
      cv_results[[counter]] <- data.frame(fold = fold, mtry = mtry_value, alpha = alpha_value, cindex = cindex)

      counter <- counter + 1
    }
  }
}

# Combine CV results
cv_results <- bind_rows(cv_results)
print(cv_results)

tuning_results <- cv_results %>%
  group_by(mtry,alpha) %>%
  summarise(mean_cindex = mean(cindex,na.rm = TRUE),
            sd_cindex = sd(cindex,na.rm = TRUE),
            min_cindex = min(cindex,na.rm = TRUE),
            max_cindex = max(cindex,na.rm = TRUE),
            .groups = "drop") %>%
  arrange(desc(mean_cindex))

print(tuning_results)

# ============================================================================
#                                 Final ORSF
# ============================================================================
# Final ORSF control based on grid search results
final_alpha <- 0.5
final_control <- orsf_control_survival(method="net",scale_x = F,ties="efron",net_mix=final_alpha)

# Fit Final ORSF Model based on grid search results
data_orsf <- data.frame(time,status,X_32062)

set.seed(42)
```

```r
final_mtry <- 20
final_orsf <- orsf(formula = Surv(time, status) ~ ., data = data_orsf, mtry = final_mtry, attach_data = T,
                   n_tree = n_tree, control = final_control, verbose_progress = T,importance = "permute")

## Variable Importance in Final ORSF Model
library(ggplot2)
library(dplyr)

# Compute the named vector of importance scores
vimp_scores <- orsf_vi_permute(final_orsf)

# Results
vimp_data <- data.frame(ID_Gene = names(vimp_scores), Importance = round(as.vector(vimp_scores), 5),
  Noise = ifelse(as.vector(vimp_scores) < 0, "Negative (Noise)", "Positive (Predictive)"))

rownames(vimp_data) <- 1:nrow(vimp_data)

# Plot horizontal barplot for genes
ggplot(vimp_data, aes(x = reorder(ID_Gene, Importance), y = Importance, fill = Noise)) +
  geom_bar(stat = "identity", width = 0.7, show.legend = TRUE) +
  coord_flip() +
  scale_fill_manual(values = c("Negative (Noise)" = "#E41A1C", "Positive (Predictive)" = "#377EB8"),
  name = "Impact Type") +
  labs(title = "Prognostic Genes by ORSF Variable Importance",x = "Genes",y = "Permutation Importance
  (OOB C-index reduction)") +
  theme_minimal(base_size = 12) +
  theme(plot.title = element_text(face = "bold", hjust = 0.5), axis.title = element_text(face = "bold"),
        legend.position = "bottom", panel.grid.minor = element_blank())

## Top 10 - Variable Selection - Backward Elimination by ORSF's Importance
top10_orsf <- orsf_vs(final_orsf, n_predictor_min=10, verbose_progress = T)

top<-c(54,2737,4055,8086,11667,12376,12509,16872,19008)

colnames(genomic_data)[top]

# ============================================================================
#                Extreme Gradient Boosting Cox Model (XGB-Cox)
# ============================================================================
# This library 'Xsurv' there is in Github: https://github.com/topycyao/Xsurv.
# This code is created for methods' paper 'Efficient gradient boosting for prognostic biomarker discovery'
library(Xsurv)

Sys.setenv(OMP_NUM_THREADS = 1) #Force single-thread execution to ensure reproducibility and stability

# Prepare data for Xsurv
x_clinical_train <- model.matrix(~ tumor_grade + grade + surgery, data = data)[, -1]
x_xgb <- data.frame(x_clinical_train,data[,-c(1:5)])
y_xgb <- data[,c(1,2)]
n <- nrow(data)

# Fit 5 - CV XGBoost model
set.seed(42)
xgb_model <- Xsurv.cv(
  datax = x_xgb,
  datay = y_xgb,
  top_n = 15,
  option = "xgb",
  method = "pl",
  nfolds = 5,
  nround = 1000,
  early_stopping_rounds = 5)

set.seed(42)
final_xgb_model <- Xsurv(
```

```
  datax = x_xgb,
  datay = y_xgb,
  top_n = 15,
  option = "xgb",
  method = "pl",
  nfolds = 5,
  nround = 1000,
  lambda = xgb_model$lambda,
  alpha = xgb_model$alpha,
  eta = xgb_model$eta)

final_xgb_model

# Final top 15 genes
top_xgb <- c(19008,19077,6149,7133,14348,3174,9465,3419,10743,4843,2737,682,11308,459,6208)
colnames(genomic_data[top_xgb])

# Plots by XGBoost Model
plot(final_xgb_model$tree$tree2) # show the interpretation tree which design the XGBoost model

plot(xgb_model$SHAP) # show the impact per top 15 variable on model output

# ============================================================================
# ============================================================================
# ============================================================================

# ------------------------------------------- End Training Code ----------------------------------------------

## ============================================================================
##                 Out of sample C - Index - Test Dataset GSE32063
## ============================================================================

# ============================================================================
#                            Preprocess Test Data
# ============================================================================
data("GSE32063_eset", package = "curatedOvarianData")
eset1 <- GSE32063_eset
genomic_data1 <- data.frame(t(exprs(eset1)),check.names = F)

pheno1 <- pData(eset1)
time1 <- pheno1$days_to_death
status1 <- ifelse(pheno1$vital_status=="living",0,1)

# Pre-processing clinical variables
pheno1$substage <- ifelse(is.na(pheno1$substage),"",pheno1$substage)
clinical_var1 <- data.frame("tumor_grade"= as.factor(paste(pheno1$tumorstage,pheno1$substage,sep="")),
                            "grade" = as.factor(pheno1$summarygrade),
                            "surgery" = as.factor(pheno1$debulking))

# Rename the names of genomic variables
GSE32063 <- data.frame("time"=time1,"status"=status1,clinical_var1,genomic_data1)
genes_name <- colnames(genomic_data1)
colnames(GSE32063)[-c(1:5)] <- paste("X",1:ncol(genomic_data1),sep = "")

# Keep the 3000 variables by MAD in training set
x1 <- GSE32063[,rownames(mad_values)[1:3000]]
test_data <- data.frame(GSE32063[,c(1:5)],scale(x1))

# ============================================================================
#                            Descriptive Analysis
# ============================================================================
# Descriptive Measures
library(psych)

summary_genomic1 <- describe(GSE32063[,-c(1:5)], quant = c(.25,.75), IQR = T, type = 2)
```

```r
hist(GSE32063$time, main = "Histogram of Overall Survival Time", xlab = "Time (in days)")
summary(GSE32063[,c(2:5)])

# ============================================================================
#             Preprocess Top 47 genes & Clinical Variables for Lasso Models
# ============================================================================
# Align categorical factor levels of the test set with the training set
test_data$tumor_grade <- factor(test_data$tumor_grade, levels = levels(data$tumor_grade))
test_data$grade       <- factor(test_data$grade, levels = levels(data$grade))
test_data$surgery     <- factor(test_data$surgery, levels = levels(data$surgery))

# Create dummy matrices for test categorical clinical factors (excluding the intercept)
x_clinical_test <- model.matrix(~ tumor_grade + grade + surgery, data = test_data)[, -1]
x_test_genes <- as.matrix(test_data[, top47])

# Combine clinical adjustments and genes to construct final test design matrix
x_test_full <- cbind(x_clinical_test, x_test_genes)

# Align columns exactly to match the training design matrix X
x_test_full <- x_test_full[, colnames(x_lasso), drop = FALSE]

# ============================================================================
#                               Cox Lasso Model
# ============================================================================
predicted_risk_lasso <- predict(cv_cox_lasso, newx = x_test_full , s = "lambda.min")

# Compute C-index
c_index_lasso <- concordance(Surv(test_data$time, test_data$status) ~ predicted_risk_lasso,reverse = T)$concordance

cat(sprintf("Linear Cox Lasso C-index: %.4f\n", c_index_lasso))

# ============================================================================
#                           Debiased Cox Lasso Model
# ============================================================================
# Predict risk scores (linear predictors) via matrix multiplication
# Applying the debiased coefficients directly to the test feature space
predicted_risk_debiased <- as.vector(x_test_full %*% b_hat_new)

# Compute Concordance Index (C-index)
y_test <- Surv(time = test_data$time, event = test_data$status)
c_index_debiased <- concordance(y_test ~ predicted_risk_debiased, reverse = TRUE)$concordance

cat(sprintf("Debiased Cox Lasso C-index: %.4f\n", c_index_debiased))

# ============================================================================
#                                 ORSF Model
# ============================================================================
# Predict survival probabilities at the median survival time
eval_time <- median(data$time)
predicted_risk_orsf <- predict(final_orsf, new_data = test_data[,c("tumor_grade","grade","surgery",top47)],
pred_horizon = eval_time)

# Compute C-index
c_index_orsf <- concordance(Surv(test_data$time, test_data$status) ~ predicted_risk_orsf,reverse = T)$concordance

cat(sprintf("Oblique Random Survival Forest (mtry = 20 & alpha = 0.5) C-index: %.4f\n", c_index_orsf))

# ============================================================================
#                               XGBoost Model
# ============================================================================
# Align categorical factor levels of the test set with the training set
test_data$tumor_grade <- factor(test_data$tumor_grade, levels = levels(data$tumor_grade))
test_data$grade       <- factor(test_data$grade, levels = levels(data$grade))
test_data$surgery     <- factor(test_data$surgery, levels = levels(data$surgery))
```

```r
# Create dummy matrices for test categorical clinical factors (excluding the intercept)
x_clinical_test <- model.matrix(~ tumor_grade + grade + surgery, data = test_data)[, -1]

# Adapt test data for predict XGBoost model
x_xgb_test <- data.frame(x_clinical_test,test_data[,-c(1:5)])
y_xgb_test <- test_data[,c(1,2)]

set.seed(42)
predicted_times <- Xsurv_predict(model = final_xgb_model$model, x_train = x_xgb, y_train = y_xgb,
x_test  = x_xgb_test)

# Compute Test Set C-index using concordance
c_index_test <- concordance(Surv(y_xgb_test$time, y_xgb_test$status) ~ predicted_times$time)$concordance

cat("Test C-index:", round(c_index_test, 4), "\n")

# ---------------------------------------------- End Testing Code ----------------------------------------------
```

Βιβλιογραφία

Salerno, S., & Li, Y. (2023). High-Dimensional Survival Analysis: Methods and Applications. *Annual Review of Statistics and Its Application* (10), 25-49.

Aghamohammadi, S. Z. (2025). Efficient and improved ridge-type shrinkage estimators in low and high dimensional cox proportional hazards regression model. *Indian Journal of Statistics* (87-B, Part 2), 400-433.

Ahn, H., & Loh, W.-Y. (1994). Tree–Structured Proportional Hazards Regression Modeling. *Biometrics* (50), 471-485.

Antoniadis, A., Fryzlewicz, P., & Letue, F. (2010). The Dantzig Selector in Cox's Proportional Hazards Model. *Scandinavian Journal of Statistics* (37), 531-552.

Barut, E., Fan, J., & Verhasselt, A. (2012). Conditional Sure Independence Screening. *Journal of the American Statistical Association*, 544-557.

Bertsimas, D., & Dunn, J. W. (2019). *Machine learning under a modern optimization lens.*

Bertsimas, D., Dunn, J., Gibson, E., & Orfanoudaki, A. (2022). Optimal survival trees. *Machine Learning* (111), 2951-3023.

Binder, H., & Schumacher, M. (2008). Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. *BMC Bioinformatics*.

Bou-Hamad, I., Larocque, D., & Ben-Ameur, H. (2011). A review of survival trees. *Statistics Surveys* (5), 44-71.

Bradic, J., Fan, J., & Jiang, J. (2011). Regularization for Cox's Proportional Hazards Model with NP-Dimensionality. *Ann Stat.* (39), 3092-3120.

Breiman, L. (2001). Random forests. *Machine Learning* (45), 5-32.

Brier, G. (1950). Verification of forecasts expressed in terms of probability. *Monthey Weather Review* (78), 1-3.

Chaturvedi, N., X. de Menezes, R., & Goeman, J. J. (2014). *Biometrical Journal* (56), 477-492.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.

Chen, Y., Jia, Z., Mercola, D., & Xie, X. (2013). A Gradient Boosting Algorithm for Survival Analysis via Direct Optimization of Concordance Index. *Computational and Mathematical Methods in Medicine*.

Chen, Y., Lam, K. F., & Liu, Z. (2024). High-dimensional feature screening for nonlinear associations with survival outcome using restricted mean survival time. *Stat, 13 (2), e673.*

Cho, H., & Hong, S.-M. (2008). Median Regression Tree for Analysis of Censored Survival Data. *Systems, Man and Cybernetics, Part A, IEEE Transactions on 38*, 715-726.

Ciampi, A., Bush, R. S., Gospodarowicz, M., & Till, J. E. (1981). An Approach to Classifying Prognostic Factors Related to Survival Experience for Non–Hodgkin's Lymphoma Patients: Based on a Series of 982 Patients: 1967–1975. *Cancer* (47), 621-627.

Ciampi, A., Chang, C. H., Hogg, S., & McKinney, S. (1987). Recursive Partition: A Versatile Method for Exploratory Data Analysis in Biostatistics. *Biostatistics*, 23-50.

Ciampi, A., Hogg, S. A., McKinney, S., & Thiffault, J. (1988). RECPAM: A Computer Program for Recursive Partition and Amalgamation for Censored Survival Data and Other Situations Frequently Occurring in Biostatistics. I. Methods and Program Features. *Computer Methods and Programs in Biomedicine* (26), 239-256.

Ciampi, A., Thiffault, J., & Sagman. (1989). RECPAM: A Computer Program for Recursive Partition and Amalgamation for Censored Survival Data and Other Situations Frequently Occurring in Biostatistics II. Applications to Data on Small Cell Carcinoma of the Lung (SCCL). *Computer Methods and Programs in Biomedicine* (30), 283-296.

Ciampi, A., Thiffault, J., Nakache, J.-P., & Asselain, B. (1986). Stratification by Stepwise Regression, Correspondance Analysis and Recursive Partition: A Comparison of Three Methods of Analysis for Survival Data with Covariates. *Computational Statistics & Data Analysis* (4), 185-204.

Collett, D. (2023). *Modelling Survival Data in Medical Research* (4th ed.). New York: Chapman & Hall (Taylor & Francis).

Dang, X., Huang, S., & Qian, X. (2021). Penalized Cox's proportional hazards model for high-dimensional survival data with grouped predictors. *Statistics and Computing*.

Dunn, J. (2018). Optimal trees for prediction and prescription. *Massachusetts Institute of Technology (PhD thesis)*.

Evers, L., & Messow, C.-M. (2008). Sparse kernel methods for high-dimensional survival data. *Bionformatics, 14* (24), 1632-1638.

Fan, J., & Gijbels, I. (1994). Censored Regression: Local Linear Approximations and Their Applications. *Journal of the American Statistical Association* (89), 560-570.

Fan, J., & Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association*, 1348-1360.

Fan, J., & Li, R. (2002). Variable Selection for Cox's Proportional Hazards Model and Frailty Model. *The Annals of Statistics* (30), 74-99.

Fan, J., & Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *Annals of Statistics* (38), 3567-2604.

Fan, J., Feng, Y., & Wu, Y. (2010). High-dimensional variable selection for Cox's proportional hazards model. *In Borrowing Strength: Theory Powering Applications – A Festschrift for Lawrence D. Brown*, 70-86.

Freund, Y., & Schapire, R. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences* (55), 119-139.

Frieadman, J., Hastie , T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* (33), 1-22.

Goeman, J. J. (2010a). L_1 Penalized Estimation in the Cox Proportional Hazards Model. *Biometrical Journal* (52), 70-84.

Gordon, L., & Olshen, R. A. (1985). Tree–structured Survival Analysis. *Cancer Treatment Reports* (69), 1065-1069.

Graf , E., Schmoor, C., Saurbrei, W., & Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine* (18), 2529-2545.

Hao, M., Lin, Y., Liu, X., & Tang, W. (2018). Robust feature screening for high - dimensional survival data. *Journal of Applied Statistics.*

Harrell Jr., F. E., Lee, K. L., & Mark, D. B. (1996). Tutorial in Biostatistics - Multivariate Prognostix Models: Issues in developing models, evaluating assumptions and adequacy, and measuring and resucing errors. *Statistics in Medicine* (15), 361-387.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982). Evaluating the yield of medical tests. *JAMA* (18), 2543-2546.

He, K., Li, Y., Zhu, J., Liu, H., Lee, J. E., Amos, C. I., et al. (2016). Component-wise gradient boosting and false discovery control in survival analysis with high-dimensional covariates. *Bioinformatics* (32), 50-57.

He, K., Wang, Y., Zhou, X., Xu, H., & Huang, C. (2019). An improved variable selection procedure for adaptive Lasso in high-dimensional survival analysis. *Lifetime Data Analysis* (25), 569-585.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* (12), 55-67.

Hong, H. G., Chen, X., Christiani, D. C., & Li, Y. (2018). Integrated Powered Density: Screening Ultrahigh Dimensional Covariates with Survival Outcomes. *Biometrics*, 421-429.

Hong, H. G., Chen, X., Kang, J., & Li, Y. (2020). The L_q-Norm Learning for Ultra-High Dimensional Survival Data: An Integrative Framework. *Stat Sin*, 1212-1233.

Hong, H. G., Kang, J., & Li, Y. (2016). Conditional screening for ultra-high dimensional covariates with survival outcomes. *Lifetime Data Anal (2018)* (24), 45-71.

Hornung, R., & Wright, M. N. (2019). Block Forests: random forests for blocks of clinical and omics covariate data. *BMC Bionformatics* (20:358), 1-17.

Hosmer David W. Jr., Lemeshow S. (1999). *Applied Survival Analysis: Regression Modeling of Time to Event Data.* Canada: John Wiley & Sons, Inc.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics, 3* (15), 651-674.

Hunter, D., & Lange, K. (2004). A tutorial on mm algorithms. *Am. Stat.*(58), 30-37.

Ishwaran, H., & Lu, M. (2019). Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival. *Stat Med.* (38), 558-582.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random Survival Forests. *The Annals of Applied Statistics, 3* (2), 841-860.

Ishwaran, H., Kogalur, U. B., Gorodeski, E. Z., Minn, A. J., & Lauer, M. S. (2010). High-dimensional variable selection for survival data. *J Am Stat Assoc* (105), 205-217.

Ishwaran, H., Kogular, U. B., Chen, X., & Minn, A. J. (2011). Random Survival Forests for High-Dimensional Data. *Statistical Analysis and Data Mining, 4*, 115-132.

Jaeger, B. C., Leann Long, D., Long, D. M., Sims, M., Szychowski, J. M., Min, Y.-I., et al. (2019). Oblique Random Survival Forests. *The Annals of Applied Statistics, 3* (13), 1847-1883.

Javanmard, A., & Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research* (15), 2869–2909.

Jin, H., Lu, Y., Stone, K., & Black, D. M. (2004). Alternative Tree–Structured Survival Analysis Based on Variance of Survival Time. *Medical Decision Making* (24), 670-680.

Kawaguchi, E. S., Suchard, M. A., Liu, Z., & Li, G. (2020). A surrogate l_0 sparse Cox's regression with applications to sparse high-dimensional massive sample size time-to-event data. *Statistics in Medicine* (39), 675-686.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *31st Conference on Neural Information Processing Systems* . Long Beach, CA, USA.

Keles, S., & Segal, M. R. (2002). Residual–Based Tree–Structured Survival Analysis. *Statistics in Medicine* (21), 313-326.

Khan, F. M., & Zubek, V. B. (2008). Support Vector Regression for Censored Data (SVRc): A Novel Tool for Survival Analysis. *Eighth IEEE International Conference on Data Mining*, (863-868).

Kim, J., Sohn, I., Sin-Ho, J., Kim, S., & Park, C. (2012). Analysis of Survival Data with Group Lasso. *Communications in Statistics - Simulation and Computation*(41), 1593-1605.

Lange, K., Hunter, D., & Yang, I. (2000). Optimization transfer using surrogate objective functions (with discussion). *J.Comput.Graph. Stat.* (9), 1-20.

Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data* (2nd ed.). Wiley-Interscience.

Leblanc, M., & Crowley, J. (1992). Relative Risk Trees for Censored Survival Data. *Biometrics* (48), 411-425.

Leblanc, M., & Crowley, J. (1993). Survival Trees by Goodness of Split. *Journal of the American Statistical Association* (88), 457-467.

Li, J., Zheng, Q., Peng, L., & Huang, Z. (2016). Survival Impact Index and Ultrahigh-Dimensional Model-Free Screening with Survival Outcomes. *Biometrics*.

Li, K., Yao, S., Zhang, Z., Cao, B., Wilson, C. M., Kalos, D., et al. (2022). Efficient gradient boosting for prognostic biomarker discovery. *Bioinformatics*, 38 (6), 1631-1638.

Lian, H., Li, J., & Tang, X. (2014). SCAD-penalized regression in additive partially linear proportional hazards models with an ultra-high-dimensional linear part. *Journal of Multivariate Analysis* (125), 50-64.

Liu, C., Liang, Y., Leung, K.-S., Chan, T.-M., Xu, Z.-B., & Zhang, H. (2013). The $L_{1/2}$ regularization method for variable selection in the Cox model. *Applied Soft Computing* (14), 498-503.

Ma, B., Yan, G., Chai, B., & Hou, X. (2022). XGBLC: an improved survival prediction model based. *Bioinformatics*, 38(2), 410-418.

Ma, Y., Li, Y., & Lin, H. (2017). Concordance Measure-based Feature Screening and Variable Selection. *Statistica Sinica* (27), 1967-1985.

Marubini, E., Morabito, A., & Valsecchi, M. G. (1983). Prognostic Factors and Risk Groups: Some Results Given by Using an Algorithm Suitable for Censored Survival Data. *Statistics in Medicine* (2), 295-303.

Mayr, A., & Schmid, M. (2013). Boosting the Concordance Index for Survival Data – A Unified Framework To Derive and Evaluate Biomarker Combinations. *PLoS ONE*, 9(1).

Mayr, A., Hofner, B., & Schmid, M. (2016). Boosting the discriminatory power of sparse survival models via optimization of the concordance index and stability selection. *BMC Bioinformatics* (17:288).

McTavish, H., Zhong, C., Achermann, R., Karimalis, I., Chen, J., Rudin, C., et al. (2022). Fast sparse decision tree optimization via reference ensembles. *In Proceedings of AAAI Conference on Artificial Intelligence.*

Mittal, S., Madigan, D., Burd, R. S., & Suchard, M. A. (2014). High-dimensional, massive sample-size Cox proportional hazards regression for survival analysis. *Biostatistics* (15), 207-221.

Molinaro, A. M., Dudoit, S., & Van Der Laan, M. J. (2004). Tree–based Statistics in Medicine Multivariate Regression and Density Estimation with Right–censored Data. *Journal of Multivariate Analysis* (90), 154-177.

Montesinos Lopez, O., Montesinos Lopez, A., & Crossa, J. (2022). *Multivariate Statistical Machine Learning Methods for Genomic Prediction.* Cham, Switzerland: Springer Nature Switzerland AG.

Park, M. Y., & Hastie, T. (2007a). L_1 - regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society Series B, Royal Statistical Society*, 69(4), 659-677.

Park, M. Y., & Hastie, T. (2007b). "glmpath: L_1 Regularization Path for Generalized Linear Models and Cox Proportional Hazards Model." R package version 0.94.

Pölsterl, S., Navad, N., & Katouzian, A. (2015). Fast Training of Support Vector Machines for Survival Analysis. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (243-259). Springer Nature.

Pölsterl, S., Navad, N., & Katouzian, A. (2016). *3rd Workshop on Machine Learning in Life Sciences.*

Rahmawati, Thamrin, S. A., Lawi, A., & Kusuma, J. (2026). Weighted Least Squares Support Vector Machine for Survival Analysis. *Statistics, Ptimization and Information Computing* (15), 243-273.

Ridgeway, G. (1999). The State of Boosting. *Computing Science and Statistics* (31), 172-181.

Sariyar, M., Schumacher, M., & Binder, H. (2014). A boosting approach for adapting the sparsity of risk prediction signatures based on different molecular levels. *Stat. Appl. Genet. Mol. Biol.*, 13(3), 343-357.

Segal, M. R. (1988). Regression Trees for Censored Data. *Biometrics* (44), 35-48.

Shivaswamy, P. K., Chu, W., & Jansche, M. (2007). A Support Vector Approach to Censored Targets. *Seventh IEEE International Conference on Data Mining*, (655-660).

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2011). Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *Journal of Statistical Software, 39* (5), 1-13.

Song, R., Lu, W., Ma, S., & Jeng, X. J. (2014). Censored rank independence screening for high-dimensional survival data. *Biometrika*, 799-814.

Steingrimsson, J. A., Diao, L., & Strawderman, R. L. (2019). Censoring Unbiased Regression Trees and Ensembles. *Journal of the American Association* (114), 370-383.

Steingrimsson, J. A., Diao, L., Molinaro, A. M., & Strawderman, R. L. (2016). Doubly robust survival trees. *Statistics in Medicine* (35), 3595-3612.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioindformatics, 307*(9).

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics, 8*(25).

Sun, H., Lin, W., Feng, R., & Li, H. (2014). Network-Regularized High-Dimensional Cox Regression for analysis of genomic data. *Stat Sin.*, 24(3), 1433–1459.

Sun, Y., Chiou, S. H., & Wang, M.-C. (2020). ROC-guided survival trees and ensembles. *Biometrics* (76), 1177-1189.

Suykens, J. K., & Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Processing Letters*, 9 (3), 293-300.

Suykens, J., Van Gestel, T., De Brabanter, J., De Moor, B., & Vandewalle, J. (2002). Least Squares Support Vector Machines. *World Scientific*. Singapore.

Therneau, T., Grambsch, P., & Fleming, T. (1990). Martingale–Based Residuals for Survival Models. *Biometrika* (77), 147-160.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society* (58), 267-288.

Tibshirani, R. J. (2009). Univariate Shrinkage in the Cox Model for High Dimensional Data. *Statistical Applications in Genetics and Molecular Biology* (8).

Tibshirani, R., & Witten, D. M. (2010). Survival analysis with high-dimensional covariates. *In Statistical Methods in Medical Research* (29-51).

Tibshirani, R., Hastle, T., Narasimhan, B., & Chu , V. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences* (99), 6567-6572.

Tibshirani, R., Simon, N., Friedman, J., & Hastie, T. (2011). Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *Journal of Statistical Software* (39).

Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., & Wei, L. (2011). On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine*, 30(10), 1105-1117.

Van Belle, V., Pelckmans, K., Suykens, J. K., & Van Huffel, S. (2009a). Additive survival least-squares support vector machines. *Statistics in Medicine* (29), 296-308.

Van Belle, V., Pelckmans, K., Suykens, J. K., & Van Huffel, S. (2009b). Feature Selection in Survival Least Squares Support Vector Machines with Maximal Variation Constraints. *Bio-Inspired Systems: Computational and Ambient Intelligence, 10th International Work-Conference on Artificial Neural Networks. Proceedings, Part I.* Salamanca, Spain.

Van Belle, V., Pelckmans, K., Van Huffel, S., & Suykens, J. A. (2011). Support vector methods for survival analysis: a comparison between ranking and regression approaches. *Artificial Intelligence in Medicine* (53), 107-118.

Verweij, P. J., & Van Houwelingen, H. C. (1994). Penalized Likelihood in Cox Regression, *Statistics in Medicine* (13), 2427-2436.

Vinzamuri, B., & Reddy, C. K. (2013). Cox Regression with Correlation based Regularization for Electronic Health Records. *IEEE 13th International Conference on Data Mining*, 758-766.

W-C, L. (1991). Survival Modeling Through Recursive Stratification. *Computational Statistics and Data Analysis* (12), 295-313.

Wang, H., & Li, G. (2019). Extreme learning machine Cox model for high - dimensional survival analysis. *Statistics in Medicine* (38), 2139-2156.

Xia, L., Nan, B., & Li, Y. (2023). Statistical inference for Cox proportional hazards models with a diverging number of covariates. *Scandinavian Journal of Statistics* (50), 550-571.

Xie, L., He, X., Keskinocak, P., & Xie, Y. (2025). Survival Analysis with Graph-Based Regularization for Predictors. *Statistics in Biosciences*.

Yang, G., Yu, Y., Li, R., & Buu, A. (2016). Feature Screening in Ultrahigh Dimensional Cox's Model. *Stat. Sin* (26), 881-901.

Yi, G. Y., He, W., & Carroll, R. J. (2021). Feature screening with large-scale and high-dimensional survival data. *Biometrics*, 1-14.

Yoshihara, K. (2012). High-Risk Ovarian Cancer Based on 126-Gene Expression Signature Is Uniquely Characterized by Downregulation of Antigen Presentation Pathway. *Clinical Cancer Research*, 18(5), 1374-1385.

Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.*, 38(2), 894-942.

Zhang, H. (1995). Splitting Criteria in Survival Trees. *In Statistical Modelling: Proceedings of the 10th International Workshop on Statistical Modeling*, (305-314).

Zhang, H., & Lu, W. (2007). Adaptive Lasso for Cox's proportional hazards model. *Biometrika* (94), 691-703.

Zhang, R., Xin, R., Seltzer, M., & Rudin, C. (2024). Optimal Sparse Survival Trees. *Proc Mach Learn Res, 238,* 352-360.

Zhao, P., & Yu, B. (2006). On Model Selection Consistency of Lasso. *Journal of Machine Learning Research* (7), 2541-2563.

Zhao, S. D., & Li, Y. (2012). Principled sure independence screening for Cox models with ultra-high-dimensional covariates. *Journal of Miltivariate Analysis* (105), 397-411.

Zhao, S. D., & Li, Y. (2014). Score Test Variable Screening. *Biometrics* (70), 862-871.

Zhou, L., Wang, H., & Xu, Q. (2016). Random rotation survival forest for high dimensional censored data. *SpringerPlus* (5:1425).

Zhou, L., Xu, Q., & Wang, H. (2015). Rotation survival forest for right censored data. *PeerJ* (3:1009).

Zhu, J., & Hastie, T. (2005). Kernel Logistic Regression and the Import Vector Machine. *Journal of Computational and Graphical Statistics* (14).

Zhu, R., & Kosorok, M. R. (2012). Recursively Imputed Survival Trees. *Journal of the American Statistical Association*, 331-340.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology* (67), 301-320.