INTERPRETABLE SPARSE MODELING OF LONGITUDINAL SIGNALS VIA CRITICAL-RANGE RECTIFICATION AND ANYTIME RULE COMPRESSION

by

Jason Orender
B.S. December 1993, University of Texas at Austin
M.S. December 2003, George Mason University
M.S. May 2018, Old Dominion University

A Dissertation Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

COMPUTER SCIENCE

OLD DOMINION UNIVERSITY
August 2026

Approved by:

Jiangwen Sun (Director)

Mohammed Zubair (Member)

Yaohang Li (Member)

Yet Nguyen (Member)

ABSTRACT

INTERPRETABLE SPARSE MODELING OF LONGITUDINAL SIGNALS VIA CRITICAL-RANGE RECTIFICATION AND ANYTIME RULE COMPRESSION

Jason Orender
Old Dominion University, 2026
Director: Dr. Jiangwen Sun

High-dimensional longitudinal data are common in clinical monitoring, industrial control systems, and other sensor-driven domains. When outcomes are governed by go/no-go thresholds and delayed effects, standard sparse models such as L1-regularized logistic regression often become unstable after lag expansion because multicollinearity undermines consistent support recovery. At the same time, many stakeholders require models that are interpretable as simple rules and that can be trained and evaluated under constrained compute. This prospectus proposes an end-to-end framework that addresses these constraints using three stages: (i) critical-range rectification, which converts each lagged continuous measurement into a {-1,+1} indicator denoting whether it lies within a data-driven range observed during events; (ii) sparse model fitting using an L1-regularized logistic baseline on the rectified design; and (iii) anytime rule compression ("logic polishing"), which collapses small-magnitude coefficients to produce a compact m-of-K rule model tuned to maximize Youden's J at a chosen operating point. Preliminary studies on synthetic data with known ground-truth rules and on real-world datasets (including the HAI industrial control system) suggest that rectification improves feature-and-lag attribution, reduces false positives, and can decrease end-to-end training time. A tractable proof-of-concept theoretical analysis uses the zero-threshold arcsin relation under joint normality to show generally how sign binarization contracts pairwise correlations, improves conditioning of the Gram matrix, and increases the likelihood of

satisfying the LASSO irrepresentable condition. The proposed dissertation work will (a) validate the proof-of-concept assertions, (b) quantify the contribution of each pipeline stage via ablations and stability studies, and (c) expand evaluation to additional longitudinal domains and baselines.

I dedicate my dissertation to my wife Danielle.

ACKNOWLEDGMENTS

I would like to thank my committee, and especially Drs. Sun and Zubair, without whose guidiance this document and the work it represents would not have been possible.

TABLE OF CONTENTS

Page

Page

Page

APPENDICES

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

Longitudinal data contain repeated measurements of the same variables across time and are central to many scientific and operational problems. Examples include symptom progression in clinical cohorts, temporal variation in biological markers, and multi-wave surveys of human decision behavior [7, 26, 32]. Across these settings, outcomes are often triggered by threshold-and-lag mechanisms: a variable enters a critical range, then a target event appears after a delay as with Figure 1. This structure is common in sensing and monitoring contexts where decision-makers need not only accurate predictions, but also clear explanations of which variables and lags drove the prediction.

In high-dimensional longitudinal modeling, a standard strategy is to construct a lag-expanded design matrix and apply a sparse estimator. This strategy is attractive because sparse models can retain predictive performance while reducing dimensionality [16, 29]. However, lag expansion also introduces strong dependence among adjacent lags and among correlated channels, creating exactly the regimes where support recovery becomes unstable and model-selection assumptions can fail [5, 36]. Practical outcomes include arbitrary swapping among correlated predictors, sensitivity to tuning choices, and weak reproducibility of selected supports across resamples.

A substantial body of work addresses these issues through penalty design and optimization advances, including elastic net, adaptive lasso, group-based penalties, and ordered lag constraints [14, 30, 35, 38, 39]. These methods provide important baselines and often improve behavior relative to plain lasso, but they still operate directly on raw correlated representations and may

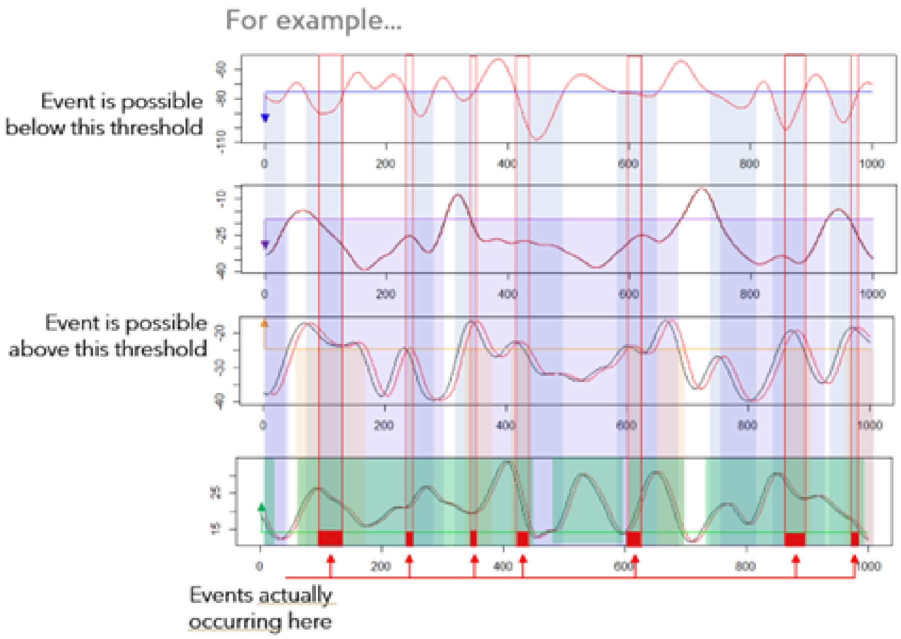

Figure 1. Illustrative threshold-and-lag behavior in longitudinal signals.

not provide explicit threshold semantics for human review. Correlation-aware variants in domain-specific contexts further reinforce this point: dependence can be mitigated, but stable, interpretable attribution remains difficult when predictors are densely coupled [10, 33, 37].

A second challenge is interpretability under high-stakes use. Post hoc explanations of complex models can be useful, but interpretable-model-first approaches are often preferable when transparency, auditability, and operational trust are mandatory [15, 22]. Rule-structured models are especially relevant because they align with human review workflows and can be validated directly by domain experts [1–3]. This dissertation therefore focuses on an inherently interpretable sparse pipeline rather than black-box prediction with after-the-fact explanation.

The proposed approach builds from prior work in this research line: a rectification-first longitudinal feature-selection method, proof-of-concept theory linking binarized transformation to improved irrepresentable-condition (IC) behavior, and an anytime rule-compression method that converts sparse models into compact logic-like forms [17–19]. The dissertation unifies these components into one framework, strengthens the supporting analysis, broadens empirical evaluation, and emphasizes reproducible implementation.

1.1 PROBLEM

This work addresses three tensions that arise jointly in longitudinal feature learning:

- Correlation + high dimension → unstable sparse support selection and weak lag attribution.
- Operational use → limited compute budgets and constraints on model complexity.
- High-stakes deployment → demand for transparent, auditable rules instead of opaque scoring functions.

The central research question is whether representation-level rectification can make sparse longitudinal selection more reliable before model fitting, and whether a principled post-selection compression step can preserve discrimination while improving human interpretability. In contrast to approaches that rely only on penalty modifications, this dissertation investigates a preprocessing-plus-selection-plus-compression pipeline designed to explicitly encode threshold behavior, reduce correlation burden, and expose causal hypotheses at the feature-lag level.

Thesis statement: critical-range rectification improves reliability of sparse longitudinal feature recovery by contracting harmful dependence structures in lag-expanded data, and anytime rule compression converts resulting sparse models into compact, operationally usable rule forms with minimal loss in discriminative performance [17, 18]. The scope emphasizes inherently interpretable sparse models with explicit attribution; deep learning appears only as contextual baselines where needed.

1.2 CONTRIBUTIONS

This dissertation makes four integrated contributions.

1. **A unified rectification-first longitudinal pipeline.** Building on prior conference work, the dissertation consolidates critical-range transformation, sparse logistic selection, and downstream rule extraction into a single end-to-end framework [18, 19].

2. **Theory-informed analysis of correlation and selection behavior.** The dissertation extends proof-of-concept arguments relating transformed representations to improved support-recovery conditions, situating these results against established lasso consistency theory and dependence-aware critiques [5, 18, 36].

3. **Anytime compression to interpretable m-of-K style rules.** Sparse coefficient models are converted into compact rule forms that can be audited and deployed under time and complexity constraints, extending recent rule-compression results [1, 17].

4. **A broad empirical protocol for longitudinal evaluation.** The dissertation compares raw and rectified representations across synthetic and real longitudinal benchmarks, including signal-rich settings such as ICS anomaly data and ionospheric radar returns, with evaluation focused on both discrimination and attribution quality [24, 25].

1.3 THESIS ORGANIZATION

Following this introduction, the dissertation is organized as follows. The background chapter formalizes longitudinal feature-learning assumptions, sparse regularization foundations, and interpretability criteria, and then positions related methods in grouped, ordered, and dependence-aware selection [6, 28, 34]. The methodology chapter presents the rectification algorithm, sparse fitting procedure, and anytime rule-compression mechanism in unified notation, including implementation details needed for reproducibility.

The experimental-design chapter defines datasets, preprocessing, temporal splits, baseline families, and statistical evaluation criteria. The results chapter reports predictive performance, support stability, lag-attribution behavior, and compression tradeoffs. The final chapters summarize limitations, future extensions, and deployment implications for longitudinal decision support. Proof details are discussed at a high level in the main text and provided step-by-step in the appendices.

CHAPTER 2

BACKGROUND

Sparse modeling with L1 penalties is a standard approach for feature selection in high-dimensional data because it often yields compact models with practical predictive performance [16, 29]. In longitudinal settings, sparse selection is also used for lag attribution: selected coefficients identify not only which variables matter, but when they matter. This chapter reviews the background needed for the dissertation, including sparse regularization, support-recovery theory under dependence, interpretability requirements, and competing method families.

Longitudinal applications motivate this framing because repeated-measure data often contain delayed and heterogeneous effects across people, systems, or environments [7, 26, 32]. When temporal predictors are expanded across lags, model dimensionality and dependence both increase, so variable selection and interpretability must be considered jointly rather than as separate post-processing steps.

2.1 L1-REGULARIZED MODELS AND FEATURE SELECTION

The LASSO and L1-regularized logistic regression induce sparsity by shrinking many coefficients to exactly zero, creating an embedded feature-selection mechanism inside model fitting [29]. This property is one reason L1 methods are widely used in longitudinal pipelines: selected nonzero lag terms can provide concise, testable hypotheses about delayed effects.

Modern implementations make these methods computationally practical at scale: coordinate-descent path algorithms with warm starts support efficient generalized linear modeling across many tuning values [6, 28]. Beyond plain L1 penalties, elastic net, adaptive lasso, and group-aware variants address correlated and block-structured predictors, which is directly relevant for lagged longitudinal features [11, 14, 35, 38, 39].

2.2 SUPPORT RECOVERY AND THE IRREPRESENTABLE CONDITION

For feature selection, prediction accuracy alone is insufficient; the key question is whether selected supports match the true active set. Zhao and Yu formalized this issue through the irrepresentable condition (IC), showing that support recovery consistency for lasso depends critically on covariance structure [36]. When off-support predictors are too correlated with active predictors, sparse recovery can fail even when predictive fit appears strong.

This limitation is central in lag-expanded longitudinal designs, where adjacent lags and related channels are frequently collinear. Reviews of lasso under dependence consistently report this selection-versus-prediction tension, with false discoveries and unstable supports common under strong correlation [5]. Correlation-aware variants, including transformed/weighted and rank-based approaches, can improve robustness in specific regimes, but they do not eliminate the underlying dependence challenge in general [20, 37].

From a linear-algebra perspective, the issue is closely tied to conditioning and inverse-Gram stability, which govern how sensitive selected supports are to perturbations in data and tuning [8]. For this reason, the dissertation treats representation-level correlation contraction as a first-class objective before sparse fitting, and uses IC-oriented reasoning as the theoretical lens for interpreting support behavior.

2.3 LONGITUDINAL LAG EXPANSION AND STRUCTURED SPARSITY

Lag expansion is the dominant way to represent temporal effects in classical supervised learning, but it can quickly create high-dimensional, redundant design matrices. Ordered and structured penalties attempt to encode temporal assumptions directly in the coefficient space, for example monotone lag decay or shared supports across tasks [30, 31].

Additional group-penalized optimization work improves runtime and practical convergence for non-orthonormal designs, making structured sparse models more usable in real pipelines [34]. These methods are important comparators for this work: they preserve the original numeric representation and optimize increasingly sophisticated penalties, whereas the proposed pipeline changes representation first and then applies sparse selection.

2.4 INTERPRETABILITY AS A DESIGN REQUIREMENT

In high-stakes settings, interpretability is not an afterthought: end-users must understand why a prediction was made and be able to verify it against domain knowledge. Interpretable-model-first arguments emphasize that transparent model classes are often preferable to post hoc explanations layered on top of black boxes [15, 22]. This principle is especially relevant when longitudinal decisions affect operations, safety, or clinical workflows.

Rule-based models provide a natural interface for such settings because they are directly inspectable and can be validated by experts testing each feature and lag individually against specific rules. Classical logical-analysis frameworks and modern optimal-rule-list methods show that transparent rule structures can be competitive while preserving auditability [1–3]. This work adopts this perspective by treating rule compression as part of model construction, not merely post-hoc explanation.

2.5 COMPETING APPROACHES

Related works will be analyzed contextually with this work in greater detail, but a brief overview of the relevant competing families include:

- Group and structured sparsity methods impose constraints on coefficients, such as groupwise inclusion or ordered lag decay; they often improve stability but generally do not alter the raw input dependence structure [14, 30, 34, 35].
- Dependence-aware sparse variants attempt to improve selection under strong correlation through reweighting, transformed objectives, or robust/rank formulations [10, 20, 37].
- Wrapper and filter selection methods remain pragmatic baselines, especially in domain pipelines where model families vary, but they are often heuristic and may provide weaker attribution consistency across resamples [12, 13].
- Robust multicollinearity-focused estimators and heuristics can improve fit under contamination and ill-conditioning, but may prioritize prediction stability over sparse interpretability [21].
- Logic-rule and LAD-style methods explicitly target threshold semantics and human-readable decisions, yet can face combinatorial scaling pressures as feature spaces grow [1–3].

Taken together, this literature motivates the design choice for this work: combine representation-level preprocessing with sparse estimation and rule-oriented compression. This places the work between two extremes, pure penalty engineering on raw lag-expanded data and direct combinatorial rule search, with the goal of balancing reliability, transparency, and computational feasibility.

CHAPTER 3

RELATED WORK

This chapter reviews method families most relevant to the problem identified in this work: sparse longitudinal feature selection under strong dependence, with interpretable downstream decision logic. The discussion is organized to move from core sparse-learning foundations to structured penalties, dependence-aware alternatives, and rule-based interpretability frameworks. The final sections position this work relative to prior work in the same method lineage.

3.1 SPARSE REGULARIZATION FOUNDATIONS

L1-regularized modeling is the canonical starting point for sparse feature selection. The lasso formulation introduced by Tibshirani established a convex mechanism that performs shrinkage and variable selection simultaneously, making it practical for high-dimensional settings [29]. In classification settings, L1-regularized logistic regression inherits the same sparsity behavior and is widely used when model compactness and feature attribution are important.

Early theoretical and empirical analyses also clarified why L1 penalties are often preferable to purely L2-regularized approaches when many predictors are irrelevant. In sparse-relevance regimes, L1 methods can scale better with growing nuisance dimensionality than rotationally invariant alternatives [16]. This intuition remains central for longitudinal lag-expanded designs, where many candidate lag terms may not be causally relevant.

A second reason this family became dominant is computational maturity. Coordinate-descent path algorithms provide efficient fitting across full regularization paths for generalized linear models and related objectives [6, 28]. These advances made sparse modeling feasible at operational scales and enabled cross-validated tuning in realistic workflows rather than one-off, expensive optimization.

However, plain lasso is not the endpoint of sparse modeling. Elastic net and adaptive lasso were proposed to improve behavior under correlation and to reduce some forms of selection bias [38, 39]. This work treats these methods as important baselines, but not as complete solutions for threshold-and-lag attribution in heavily collinear longitudinal representations.

3.2 SUPPORT RECOVERY UNDER DEPENDENCE

Predictive accuracy and correct feature recovery are distinct goals. Zhao and Yu formalized this distinction by showing that lasso model-selection consistency depends on covariance-structure conditions, particularly the irrepresentable condition (IC) [36]. In practical terms, when inactive predictors are too correlated with active predictors, support recovery may fail even if prediction remains competitive.

This issue is acute in lag-expanded longitudinal data, where neighboring lags and related channels often have strong dependence. Recent reviews emphasize that lasso and many derivatives can produce unstable supports and elevated false positives in dependent designs, especially under common tuning choices optimized for prediction rather than exact support recovery [5].

Dependence-aware variants address parts of this problem. Examples include transformed or weighted formulations for highly correlated predictors and robust rank-based alternatives that reduce sensitivity to link-function misspecification or heavy-tailed behavior [20, 37]. These methods improve robustness in specific regimes, but they generally do not provide a unified threshold semantics and lag-rule representation for operational interpretation.

From a matrix-analysis viewpoint, these selection instabilities reflect conditioning and inverse-Gram sensitivity, which directly affect sparse-support perturbation behavior [8]. This perspective motivates representation-level interventions before sparse fitting, rather than relying only on penalty redesign in the original feature space.

3.3 GROUP, BLOCK, AND ORDERED SPARSITY

Structured penalties were developed to stabilize selection when predictors have known organization. Early simultaneous-selection formulations for multiresponse models and later grouped penalties introduced mechanisms that select predictors at the block level rather than coefficient-by-coefficient [31, 35]. This is relevant in longitudinal settings where features can be grouped by channel, basis expansion, or lag family.

Subsequent work extended grouped sparsity to logistic settings and improved practical optimization. Group lasso for logistic regression, blockwise sparse regression, and unified majorization-descent solvers significantly improved feasibility for high-dimensional non-orthonormal designs [11, 14, 34]. These methods remain strong comparators because they encode structural assumptions directly in the objective while retaining convexity in many cases.

Temporal structure has also been added explicitly through ordered penalties. Ordered lasso introduces monotonicity constraints across lag coefficients, reflecting assumptions such as decaying lag influence and yielding interpretable lag profiles when those assumptions hold [30]. This is a particularly relevant baseline for relevant experiments because it targets time-lagged sparsity directly.

Despite these advances, structured penalties still operate in the original correlated representation. They can improve stability and interpretability relative to plain lasso, but they do not inherently convert continuous trajectories into explicit threshold logic. This work positions its rectification step as complementary to these methods: change the representation first, then apply sparse learning.

3.4 CORRELATION-AWARE AND ALTERNATIVE FEATURE SELECTION FAMILIES

Beyond penalized linear models, several approaches attempt to reduce redundancy through direct relevance-redundancy criteria. Katrutsa and Strijov proposed a quadratic-programming feature-selection framework where pairwise similarity and target relevance are balanced in a convex relaxation [10]. This strategy can reduce redundant supports, but it is primarily pairwise in construction and does not natively express lag-threshold rules.

Robust multicollinearity-focused estimators provide another pathway. Methods combining robust regression ideas with conditioning-oriented penalties can improve estimation under contamination and ill-conditioning [21]. These approaches are useful for robustness analysis but typically prioritize fit stability over compact, human-auditable sparse logic.

Filter, wrapper, and embedded selection families remain widely used in practice because they are flexible and model-agnostic [13]. Recent hybrid ranking-plus-search approaches also show good empirical performance in domain-specific pipelines [12]. Still, these families can be sensitive to search heuristics, may yield unstable subsets under resampling, and usually do not produce direct temporal-threshold semantics without additional modeling layers.

Accordingly, this work includes these methods primarily as conceptual comparators rather than as primary interpretability baselines. Their strengths are pragmatic feature screening and broad applicability; their limitations are weaker guarantees for sparse lag attribution and logic-level interpretability.

3.5 INTERPRETABLE MODELING AND RULE-LEARNING LITERATURE

Interpretability-focused literature increasingly argues that high-stakes deployments should prefer inherently interpretable models over post hoc explanations of black boxes whenever performance is competitive [15, 22]. This framing strongly influences the design choices in this work: interpretability is treated as a primary objective, not a secondary (post-hoc) reporting artifact.

Rule-based models are central to this viewpoint. Classical Logical Analysis of Data (LAD) frameworks formalize binarization and logic-pattern extraction, including optimization-based treatment of cut-point selection and pattern construction [2, 3]. These methods establish the value of explicit threshold logic for decision support, but can face combinatorial scaling limits in large, highly correlated feature spaces.

Modern rule-list methods demonstrate that transparent models can be both accurate and principled when optimization is carefully designed. Certifiably optimal rule-list learning shows that exact or bounded-search approaches can produce compact, auditable models with competitive predictive performance on structured tasks [1]. This work does not replicate global rule-list optimization, but it draws from the same interpretability objective and complexity-aware evaluation philosophy.

Relative to this literature, the strategy for this work is to anchor interpretability in sparse model fitting and then compress coefficients into logic-like forms through an "anytime" procedure. This avoids full combinatorial search while still yielding compact rule structures suitable for expert review.

3.6 LONGITUDINAL HIGH-DIMENSIONAL MODELING CONTEXT

Longitudinal methods in broader statistics and machine learning emphasize that repeated measures introduce heterogeneity, temporal dependence, and subgroup structure that can invalidate cross-sectional assumptions. Model-based clustering with regularized mixed-effects formulations demonstrates one advanced path for high-dimensional trajectory modeling [33]. These methods are powerful for discovering latent trajectory classes, but they optimize a different end goal than sparse, explicit lag-rule extraction.

Domain literature further illustrates why interpretable longitudinal modeling matters. Clinical and biological longitudinal studies report heterogeneous temporal pathways and delayed outcome behavior, reinforcing the need for methods that preserve temporal attribution [7, 32]. Similar heterogeneity appears in socio-environmental longitudinal surveys [26]. These contexts motivate models that can be inspected and discussed by domain experts rather than only ranked by black-box metrics.

Benchmark infrastructure also matters for evaluating longitudinal methods. Public anomaly and signal datasets used in this line of inquiry, such as HAI ICS telemetry and ionospheric radar-return data, provide realistic stress tests for lag attribution, sparse recovery, and interpretability tradeoffs [24, 25]. Related engineering contexts with threshold-driven transitions likewise support the relevance of range-based temporal reasoning [9].

3.7 METHOD LINEAGE AND DISSERTATION POSITIONING

This work is not a stand-alone methodology detached from prior work; it extends a concrete sequence of methods developed in earlier publications. The initial conference work introduced a rectification-first perspective for longitudinal feature learning, where critical-range binarization precedes sparse selection [19]. This established the basic preprocessing-plus-lasso pipeline and empirical motivation.

Subsequent work added proof-of-concept theory and broader case studies. The 2025 efficient longitudinal feature-selection paper linked binarized transformation to correlation contraction arguments and improved IC-related behavior under stated assumptions, while also reporting computationally practical implementations [18]. This paper provides the main theoretical bridge between representation design and support recovery.

The companion 2025 anytime rule-compression paper addressed the remaining interpretability gap: sparse coefficients are often still too numerous for direct human use. The anytime compression stage converts sparse rectified models into compact m-of-K style rule behavior with controlled discrimination tradeoffs and early-stop practicality [17]. This contribution aligns the pipeline with interpretable-model deployment requirements discussed earlier in this chapter.

Compared with related work, the dissertation's distinguishing emphasis is integration. Rather than proposing only a new penalty, only a new binarizer, or only a new rule learner, it unifies all three stages and evaluates them as one longitudinal decision-support pipeline. The key comparative hypothesis is that representation-level rectification plus sparse selection plus anytime compression better balances support reliability, interpretability, and compute constraints than any single-stage alternative.

3.8 SUMMARY OF GAPS IN PRIOR WORK

The reviewed literature reveals four persistent gaps that motivate this dissertation.

- Sparse penalties are computationally mature and often predictive, but support consistency remains fragile under strong lag-induced dependence [5, 36].

- Structured and dependence-aware penalties improve behavior in specific regimes, yet usually remain tied to raw correlated representations and do not produce threshold-native logic outputs [30, 35, 37].

- Rule-learning frameworks deliver transparency, but direct combinatorial optimization can be difficult to scale in high-dimensional longitudinal spaces [1, 2].

- Prior work in this dissertation line introduced the key pieces, but an end-to-end, reproducible, and extensively benchmarked synthesis is still needed [17–19].

The remainder of the dissertation addresses these gaps through unified methodology, theory-informed analysis, and comparative empirical evaluation.

CHAPTER 4

RESEARCH QUESTIONS

This chapter defines the core research questions that guide this work and explains why they are important. Longitudinal feature learning is targeted under three simultaneous constraints: strong dependence created by lag expansion, the need for reliable feature-lag attribution, and the need for compact human-auditable model outputs. Prior chapters established that sparse methods are attractive but not always stable under dependence, and that interpretability requirements in high-stakes settings favor inherently transparent model forms over post hoc explanation layers [5, 15, 22, 29, 36].

The three research questions are:

- **RQ1:** Can critical-range rectification produce a stable sparse baseline with reliable feature and lag attribution?
- **RQ2:** Why does rectification work when it works, and how far can that mechanism be generalized beyond current proof-of-concept assumptions?
- **RQ3:** Can the resulting sparse solution be compressed into compact logical rules without unacceptable loss in operational discrimination?

Together, these questions move from empirical utility (**RQ1**) to theoretical credibility (**RQ2**) to deployment usability (**RQ3**). They are intentionally sequenced so that later questions build on earlier ones: there is little value in rule compression if upstream sparse attribution is unstable, and little scientific value in observed gains if the mechanism is not understood.

4.1 WHY RQ1 MATTERS

RQ1 addresses the practical entry point of the dissertation: does representation-level rectification make sparse longitudinal selection more reliable than fitting directly on raw lag-expanded inputs? This question is important because dependence among lagged predictors is the norm in longitudinal data, and support recovery can degrade sharply in such settings even when prediction remains acceptable [5, 36].

Competing methods attempt to mitigate this problem through penalty design (elastic net, adaptive lasso, group/ordered penalties) or relevance-redundancy optimization [10, 14, 30, 35, 38, 39]. These methods are essential baselines, but they still work primarily in the original correlated representation. RQ1 matters because it tests a different intervention point: transform representation first, then apply sparse learning.

This question is also important for continuity with prior publications in the dissertation line. Earlier work introduced and refined rectification-first sparse longitudinal modeling and reported promising empirical behavior across synthetic and real datasets [18, 19]. This work must now determine whether those improvements are robust across broader benchmark designs, stronger baselines, and stricter attribution-oriented criteria.

4.2 WHY RQ2 MATTERS

RQ2 asks for mechanism, not only outcome. Even if rectification improves support stability in multiple datasets, this work has taken upon itself to explain *why* that improvement occurs and under what assumptions it should be expected. This matters for scientific defensibility, external validity, and principled method extension.

At a high level, the motivating hypothesis is that rectification can contract harmful dependence structure in ways that improve sparse-support conditions related to IC-style reasoning [8, 36]. Prior proof-of-concept analysis in this line supports this direction under explicit assumptions, but does not yet establish universal guarantees across all threshold settings and dependence regimes [18]. RQ2 is therefore critical: it separates durable methodological insight from dataset-specific heuristic success.

RQ2 also matters for how results will be interpreted in later chapters. If gains appear only in certain dependence geometries or event-logic structures, that boundary must be made explicit. A clear mechanism and boundary analysis prevents overclaiming and supports reproducible, domain-appropriate deployment.

4.3 WHY RQ3 MATTERS

RQ3 addresses the final translational step: can sparse rectified models be converted into compact logical rules that humans can audit and use? In high-stakes settings, model acceptance often depends on transparency, verifiability, and ease of expert review, not only ROC-level discrimination [15, 22].

Rule-oriented model families and LAD-style traditions demonstrate the value of explicit logic structures for decision support [1–3]. However, direct combinatorial search can be expensive in high-dimensional longitudinal spaces. The dissertation's question is whether an anytime compression stage can deliver much of the same interpretability benefit while preserving most of the discriminative value of the sparse baseline [17].

RQ3 is important because it determines operational usability. A model with good AUC but hundreds of small coefficients is often difficult to deploy in environments that require policy documentation, cross-team review, and traceable rationale. Compact rules can close that gap if their performance loss is controlled and measurable.

4.4 CROSS-RQ EVALUATION PRIORITIES

Detailed experimental designs are reserved for later chapters, but the dissertation applies a common set of evaluation priorities across all three research questions:

- **Predictive discrimination:** metrics such as AUC and Youden's J at operational thresholds, with context-dependent interpretation across datasets [18, 24, 25].
- **Attribution quality and stability:** support recovery behavior, lag localization fidelity, and stability across folds or resamples.
- **Interpretability complexity:** active-feature count, rule count, and rule length as direct proxies for human audit burden [1, 15].
- **Efficiency:** end-to-end training and inference time, including preprocessing overhead and compression overhead.
- **Practical non-inferiority framing where appropriate:** when simplified rule models are compared to upstream sparse baselines, interval-based equivalence logic can be used to assess whether differences are practically small [23].

These priorities reflect the overall objective: not maximizing a single metric, but balancing reliability, interpretability, and computational feasibility.

4.5 SIGNIFICANCE OF THE CHAPTER

This chapter establishes the scope and rationale of the research program before technical detail. RQ1 asks whether the approach works empirically in the intended problem class. RQ2 asks whether the observed behavior is theoretically credible and generalizable. RQ3 asks whether the resulting models are usable in real decision workflows. Taken together, the three questions define the dissertation's central claim that rectification-first sparse longitudinal learning, followed by anytime rule compression, can provide a practical middle ground between unstable raw sparse fitting and computationally heavy direct rule search [17–19].

CHAPTER 5

PRELIMINARY STUDIES AND EVIDENCE TO DATE FOR RQ1

RQ1 asks whether critical-range rectification can produce a stable sparse baseline with reliable feature and lag attribution in longitudinal settings. This chapter summarizes the current evidence accumulated before the full dissertation-scale evaluation. The goal here is not to present the final, exhaustive answer, but to establish what has already been demonstrated, what remains uncertain, and why the observed results are important for the broader research program.

The evidence reported to date is anchored in prior work from the same method lineage, beginning with the rectification-first lasso logic framework and extending to later theory-backed case studies [18, 19]. Across studies, the central empirical pattern is consistent: when the data generating process has threshold-and-lag structure, rectification tends to improve support concentration, lag localization, and sparse-model usability under multicollinearity pressure.

5.1 HOW RQ1 IS EVALUATED IN PRELIMINARY WORK

To answer RQ1 in a meaningful way, preliminary studies use three complementary evidence streams:

- **Controlled synthetic studies** with known feature-lag ground truth, where attribution quality can be inspected directly.
- **Comparative baseline studies** against methods designed to address collinearity or redundancy through alternative mechanisms.

- **Real-world case studies** that test whether observed gains survive outside synthetic conditions.

The preliminary decision criteria are practical rather than purely theoretical: discriminatory performance (for example AUC and Youden's J), sparsity and support clarity, lag-level attribution fidelity, and end-to-end computational cost. This aligns with known limitations of penalty-only sparse methods under dependence, where prediction may remain strong while support recovery becomes unstable [5, 36].

5.2 SYNTHETIC EVIDENCE WITH KNOWN GROUND TRUTH

Synthetic longitudinal datasets are constructed so that events occur when a small subset of variables simultaneously enters critical ranges at specific lags. This creates a hard but controlled setting where lag expansion induces severe multicollinearity by design, closely mirroring the failure modes discussed in lasso consistency literature [5, 36]. In this regime, representation-level rectification is applied before sparse fitting and compared with untransformed baselines.

The synthetic results in Table 1 show the main preliminary pattern: transformed models improve discrimination and often reduce runtime relative to their untransformed counterparts. More importantly for RQ1, support concentration is visibly improved at the coefficient level when transformed inputs are used, which directly affects lag-attribution reliability.

Figure 2 provides qualitative support for the same conclusion: rectification yields nonzero coefficients that align more closely with true causal lags, while untransformed fitting produces noisier support spread. This behavior is consistent with the rectification rationale described in prior publications, where binarized critical-range mapping is expected to reduce harmful correlation effects before sparse optimization [18, 19].

Table 1. Synthetic case: transformed vs untransformed comparisons (values from preliminary experiments).

Method	Transform Status	Rel. Time	Youden's J	F1
LASSO	Transformed	1.00	0.975	0.988
LASSO	Untransformed	1.83	0.753	0.804
Random Forest	Transformed	12.5	0.980	0.990
Random Forest	Untransformed	43.0	0.913	0.947
Group LASSO	Transformed	24.5	0.689	0.558
Group LASSO	Untransformed	53.5	0.624	0.459

5.3 COMPARISON TO COMPETING FEATURE-SELECTION BASELINES

Preliminary RQ1 evidence also includes comparisons with methods intended to address redundancy or grouped dependence through different mechanisms. Group and block-structured sparse families remain important comparators because they are explicitly designed for correlated predictor settings [11, 14, 30, 34, 35]. In these early studies, however, representation-level rectification combined with standard lasso often provides clearer lag attribution in threshold-triggered synthetic settings.

A focused comparator is the quadratic-programming selector of Katrutsa and Strijov, which optimizes a relevance-redundancy objective using pairwise similarity structure [10]. The method is conceptually aligned with the goal of reducing redundancy, but it emphasizes pairwise criteria and does not directly encode multi-lag conjunctive trigger logic.

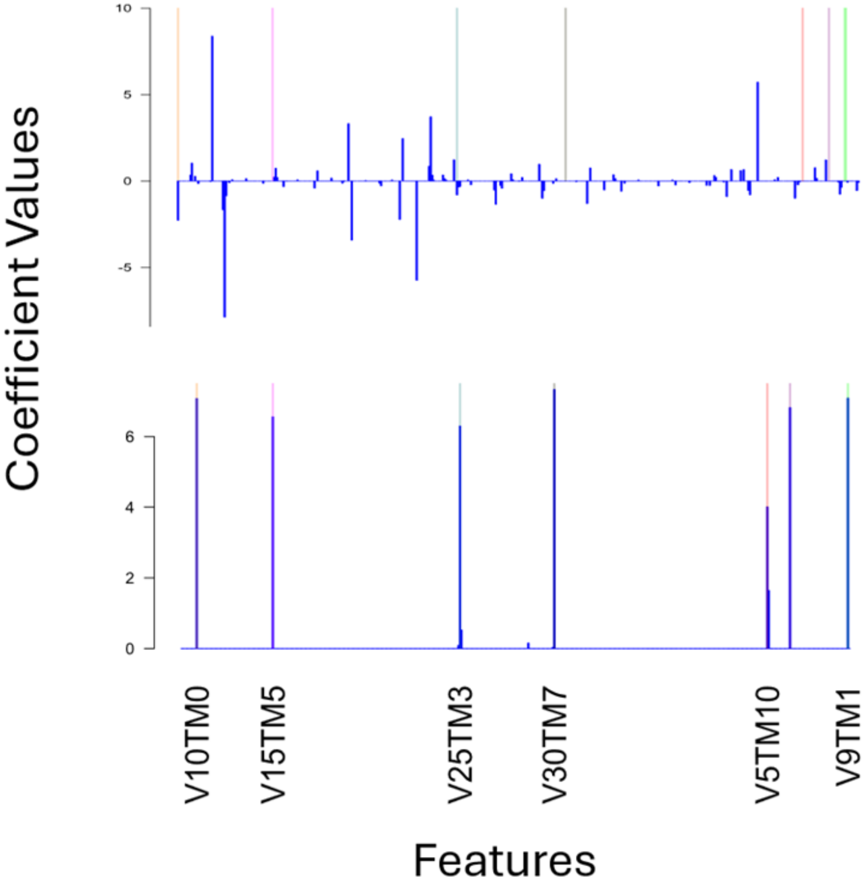

Figure 2. Coefficient profiles on synthetic data: transformed vs untransformed (ground-truth lag locations highlighted).

In preliminary tests, the quadratic-programming approach often identifies related variables but may miss the correct lag localization and exhibits substantially higher computational cost. Observed runtime differences were on the order of roughly 100–200x relative to rectification plus lasso in the tested configuration. For RQ1, this matters because a "stable sparse baseline" is not only a statistical target; it must also be practical to run repeatedly across folds, ablations, and deployment refresh cycles.

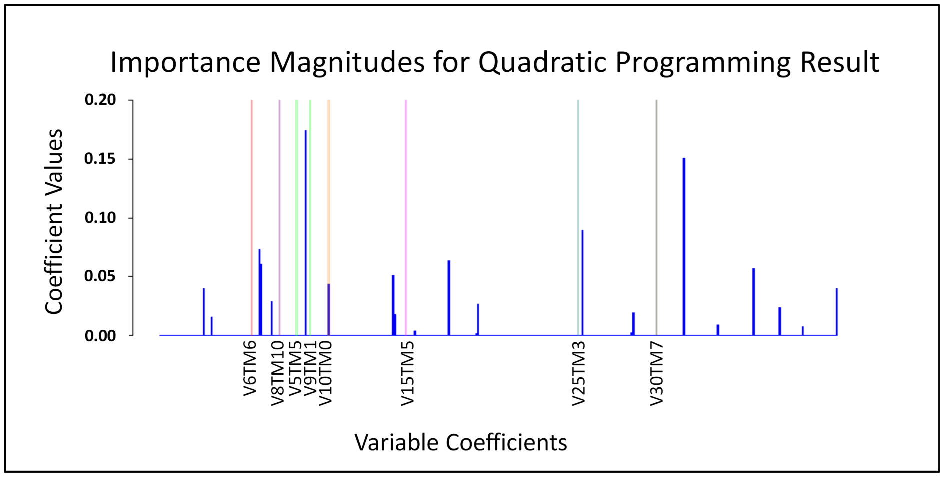

Figure 3. Quadratic-programming importance magnitudes on the synthetic dataset.

5.4 REAL-WORLD EVIDENCE AND INTERPRETABILITY TRADEOFFS

Synthetic evidence is necessary but not sufficient. Preliminary real-world studies therefore evaluate whether the same pattern appears in operationally relevant longitudinal data. The HAI industrial control benchmark is particularly useful because it includes realistic multi-sensor temporal dynamics with labeled attack scenarios and known subsystem context [24]. Additional evidence from historical ionospheric radar work supports the relevance of lagged signal discrimination settings where sparse attribution can complement raw predictive performance [25].

As shown in Table 2, HAI results favor transformed models on both discrimination and attribution-oriented interpretation. In this setting, transformed models concentrate coefficients on turbine-related channels aligned with known attack context, while untransformed models spread attribution more diffusely.

Cross-dataset behavior is not uniform. In the UNICEF case, untransformed models can achieve stronger raw discrimination metrics while transformed models remain substantially sparser and

Table 2. Real-world case studies: transformed vs untransformed comparisons (values from preliminary experiments).

Statistic	HAI (Transformed)	HAI (Un-Trans)	UNICEF (Transformed)	UNICEF (Un-Trans)
ACC	0.987	0.939	0.902	0.963
AUC	0.982	0.943	0.865	0.989
F1	0.987	0.942	0.894	0.958
Youden's J-Index	0.974	0.878	0.815	0.925

easier to interpret. This pattern is consistent with prior reports that rectification benefits are context dependent: strongest in threshold-and-lag aligned regimes, and more mixed when data-generating structure is less compatible with the assumed critical-range mechanism [18].

5.5 INTERIM ANSWER TO RQ1

Based on evidence to date, the preliminary answer to RQ1 is **yes, conditionally**. Critical-range rectification can produce a more stable and interpretable sparse baseline, particularly in settings where events are driven by threshold-triggered lag interactions and raw lag expansion creates severe dependence. Evidence supporting this claim includes:

- improved lag-attribution concentration in controlled synthetic experiments,
- favorable or comparable discrimination with better sparsity in key real-world cases,
- and materially lower runtime than certain redundancy-oriented alternatives in tested config-

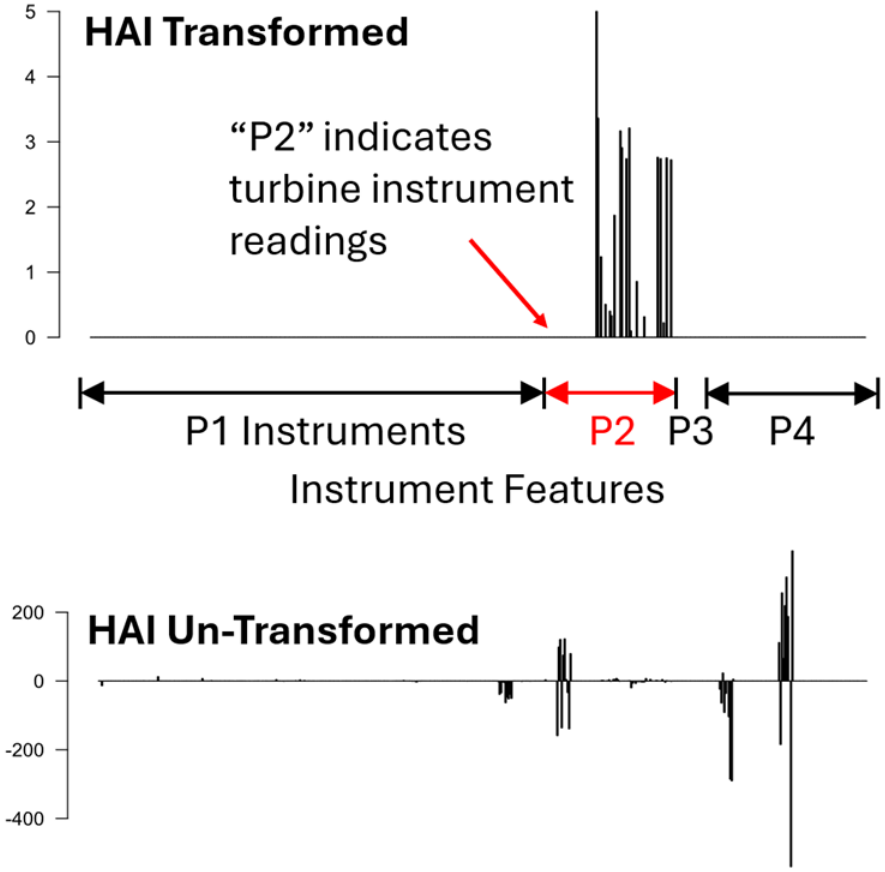

Figure 4. Real-world coefficient concentration and performance: HAI turbine-loop selection vs untransformed spread.

urations.

However, the current evidence is not yet sufficient for universal claims. Remaining gaps include broader dataset diversity, stronger robustness checks for edge-case correlation structures, and tighter statistical characterization of when discrimination gains versus interpretability gains dominate. These limitations motivate the deeper analyses in subsequent chapters, where RQ1 is tested under expanded protocols and connected to the theoretical questions addressed by RQ2.

CHAPTER 6

THEORETICAL ANALYSIS (RQ2)

RQ2 asks why rectification helps and how broadly that mechanism can be generalized. This chapter provides a proof-of-concept (PoC) theoretical justification for the binarization step using the irrepresentable condition (IC) as the central lens. The analysis is intentionally scoped: it focuses on zero-threshold sign binarization under joint normal feature assumptions, where correlation behavior is analytically tractable [18, 36].

The point of this chapter is not to claim a universal theorem for every thresholding scheme. Instead, it establishes a defensible mechanism: if binarization contracts dependence in the right way, the IC becomes easier to satisfy, and sparse support recovery becomes more likely. That mechanism aligns with both prior empirical evidence in this line of inquiry and broader observations that lasso-style selection can be unstable under strong dependence [5, 18].

6.1 WHY RQ2 MATTERS AFTER RQ1 EVIDENCE

RQ1 showed encouraging preliminary evidence that rectification improves support concentration and lag attribution in several settings. However, empirical success alone is insufficient for the central claim. Without a credible mechanism, improvements could be dismissed as dataset-specific tuning effects.

For sparse selection methods, this concern is especially important. Lasso is a strong predictive tool [29], but model-selection consistency depends on structural conditions that may fail under multicollinearity [36]. In other words, high AUC does not imply reliable feature recovery. RQ2 therefore asks whether rectification improves a known structural bottleneck rather than merely shifting surface metrics.

6.2 IC IN COMPUTER-SCIENCE TERMS

The IC can be read as an interference constraint in a sparse-recovery system. Let S be the true active-feature set and S^c the inactive set (the complement). For lasso to recover the right support, inactive features should not be too well explained by active features after accounting for active-feature geometry. Formally, one common sufficient condition is:

$$\left\| \mathbf{X}_{S^c}^{\top} \mathbf{X}_S (\mathbf{X}_S^{\top} \mathbf{X}_S)^{-1} \operatorname{sign}(\beta_S) \right\|_{\infty} < 1.$$

A CS-oriented interpretation is useful:

- $\mathbf{X}_S^{\top} \mathbf{X}_S$ is the active-feature interaction matrix.
- $(\mathbf{X}_S^{\top} \mathbf{X}_S)^{-1}$ is a de-mixing operator for that interaction.
- $\mathbf{X}_{S^c}^{\top} \mathbf{X}_S$ measures leakage from active features into each inactive feature.
- The ℓ_∞ bound asks for worst-case leakage to stay below a hard margin.

When this leakage margin is violated, inactive features can mimic active ones and enter the model, taking the place of true support in the model, causing feature and lag misattribution due to spurious correlation. This is a structural reason for false inclusions and support instability, consistent with dependency-focused reviews [5].

6.3 POC SETUP AND ASSUMPTIONS

To keep the analysis explicit, this chapter uses the PoC setting from prior work [18].

Data and transformation. Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be standardized continuous features and $\tilde{\mathbf{X}} \in \{\pm 1\}^{n \times d}$ the sign-binarized version:

$$\tilde{x}_{ij} = \operatorname{sign}(x_{ij}),$$

with threshold at zero. For derivations, the appendix sometimes uses an equivalent $\{0,1\}$ encoding; after centering, it preserves the same correlation-ordering arguments up to constant scaling.

Assumptions.

- Continuous features are jointly normal after standardization.
- The support set S is fixed in the local analysis.
- Correlation summaries are interpreted through large-sample covariance approximations.
- The goal is comparative: raw versus binarized IC satisfaction probability.
- The zero-threshold is meaningful and retains the signal components relevant to the PoC mechanism in the target regime (e.g. modeling a threshold-based phenomena).

Joint normality and zero-threshold meaningfulness are modeling assumptions introduced for tractable PoC analysis; they are not empirically verified prerequisites for using the full pipeline.

These assumptions are restrictive by design. They allow closed-form dependence mapping and a clean theorem statement, but they are not presented as full generality. In finite samples, especially when d is large relative to n, population-covariance approximations may deviate from realized sample behavior, so the claims here are regime-scoped and probabilistic.

6.4 WHY SIGN BINARIZATION CAN REDUCE DEPENDENCE

Under bivariate normality, sign-binarized correlation is a deterministic function of raw Pearson correlation (ρ):

$$\tilde{\rho} = \frac{2}{\pi}\arcsin(\rho).$$

For $\rho \in (-1,1)$, this mapping contracts magnitude in most of the interior, while preserving sign and endpoint behavior. Intuitively, sign mapping removes amplitude information and retains only direction, which can weaken linear coupling caused by shared scale variation. All binarized equivalent values will be annotated with a tilde (e.g. "$\tilde{\rho}$" for the binarized Pearson correlation).

From an IC viewpoint, this contraction can help in two places:

- It can reduce cross-correlation between inactive and active features.

- It can improve conditioning behavior of the active Gram block in relevant regimes.

The second point should not be read as a universal monotone guarantee for every design. It is a regime-level effect supported in the PoC analysis and tied to matrix-conditioning arguments [8, 18].

6.5 LEMMA PATHWAY (WITH INTUITION)

This section summarizes the proof structure. Full derivations are moved to the appendix.

6.5.1 Lemma 1: Correlation Contraction After Binarization

$$\tilde{\rho} = \frac{2}{\pi}\arcsin(\rho).$$

Interpretation: Under the PoC assumptions, sign binarization induces a deterministic arcsine mapping from raw correlation to transformed correlation. In practical terms, the representation keeps polarity while suppressing magnitude information that drives linear coupling.

6.5.2 Lemma 2: Controlled Inverse Behavior for Active Block

$$|\tilde{\rho}| \leq |\rho|.$$

Interpretation: The transformed correlation magnitude is bounded by the raw magnitude (with equality at endpoints). This bound is the key contraction property used downstream to control active-block inversion effects and reduce amplification risk in de-mixing.

6.5.3 Lemma 3: Lower Inactive-Active Leakage

$$\left|\mathrm{Cov}(\tilde{x}_{S^c i}, \tilde{x}_{S j})\right| \leq \left|\mathrm{Cov}(x_{S^c i}, x_{S j})\right|.$$

Interpretation: For each inactive-active pair, binarization weakens covariance leakage (or leaves it unchanged in edge cases). In sparse-selection terms, inactive features become less able to mimic the active span.

6.5.4 Lemma 4: Easier IC Satisfaction

$$\|\tilde{G}^{-1}\|_\infty \leq \|G^{-1}\|_\infty \quad \text{(PoC positive-correlation regime).}$$

Interpretation: The transformed active block requires no greater inverse amplification than the raw block in the target regime. Combined with Lemma 3, this yields a smaller IC interference term and therefore a higher chance of satisfying the IC margin.

6.6 MAIN THEOREM (PROBABILITY FORM)

Let

$$\Theta(\mathbf{X}) = \left\| \mathbf{X}_{S^c}^{\top} \mathbf{X}_S (\mathbf{X}_S^{\top} \mathbf{X}_S)^{-1} \operatorname{sign}(\beta_S) \right\|_{\infty},$$

and define $\Theta(\tilde{\mathbf{X}})$ analogously for binarized features. In the PoC setting, the target inequality is

$$P\big(\Theta(\mathbf{X}) < 1\big) \leq P\big(\Theta(\tilde{\mathbf{X}}) < 1\big).$$

This theorem states a probabilistic improvement claim, not deterministic per-instance dominance. The claim is that transformed designs are at least as likely to lie in an IC-favorable region. Strict improvement is expected only in stylized regimes.

6.7 PROOF SKETCH

A concise proof sketch is given below; full derivations remain in the appendix.

Sketch. For each inactive feature index $i \in S^c$, define

$$\theta_i = \left| \mathbf{c}_i \mathbf{G}^{-1} \operatorname{sign}(\beta_S) \right|,$$

where $\mathbf{c}_i$ is its covariance vector with active features and $\mathbf{G}$ is the active Gram matrix.

In large-sample approximation, replace sample quantities by population correlation blocks. For raw and binarized designs this gives comparable expressions based on (ρ_{iS}, ρ_{SS}) and $(\tilde{\rho}_{iS}, \tilde{\rho}_{SS})$ respectively.

By Lemma 1, the transformed inactive-active block contracts in norm under the PoC regime. By Lemma 2, inverse amplification from the active block is controlled. Using submultiplicative norm bounds,

$$\theta_i^{\text{bin}} \lesssim \|\tilde{\rho}_{iS}\| \, \|\tilde{\rho}_{SS}^{-1}\| \, \|\operatorname{sign}(\beta_S)\|,$$

with an analogous bound for θ_i in the raw domain. The transformed upper bound is tighter under the lemma conditions, which implies a weakly higher probability that each inactive-feature interference score stays below 1. Taking the maximum over $i \in S^c$ yields the stated IC-probability improvement. □

6.8 PRACTICAL INTERPRETATION

The theorem can be read as a data-representation effect on sparse search geometry.

- **Before transformation:** lag-expanded features create dense dependency graphs. Multiple inactive nodes have high edge weight to active nodes, increasing support ambiguity.
- **After transformation:** edge weights are attenuated in the target regime, reducing ambiguity and lowering the number of near-tie candidates along the regularization path.

This interpretation connects theory to observed outcomes in earlier chapters: fewer false positives, clearer lag attribution, and higher support stability in threshold-and-lag aligned datasets [18, 19].

Importantly, the argument does *not* require modifying lasso internals. The solver, objective, and tuning workflow remain standard; the intervention is in representation. That separation is operationally useful because it preserves mature optimization tooling while changing the statistical geometry presented to the solver.

6.9 A SMALL INTERFERENCE EXAMPLE

Consider a simplified setting with one active feature x_1 and one inactive feature x_2 that is strongly correlated with x_1. If the true signal is carried only by x_1, lasso can still assign weight to

x_2 when x_2 is highly predictable from x_1. In IC language, this is a high-leakage case: x_2 sits close to the active span, so the inactive-feature margin shrinks.

Now apply sign binarization. In the PoC regime, the effective correlation between the two channels is contracted by the arcsin mapping. The inactive channel remains related to the active channel, but less linearly confounded in the space used by the sparse solver. This can move the system from a near-violation regime toward a satisfied-margin regime.

For an advanced CS reader, this is analogous to reducing feature aliasing in a compressed representation. The model family is unchanged, but representational overlap among candidate predictors is reduced enough that greedy path updates and KKT-driven screening are less likely to elevate spurious coordinates early in the path.

6.10 HOW THIS DIFFERS FROM PENALTY-ONLY FIXES

A natural question is why this chapter emphasizes representation rather than only stronger penalties. Existing work already proposes many penalty-level fixes for dependence, including elastic net, adaptive penalties, and grouped or ordered structures [30, 35, 38, 39]. These methods are important and remain part of the baseline set.

The RQ2 claim is narrower: even with the same downstream sparse learner, changing the representation can improve the structural conditions that drive support recovery. Penalty redesign and representation redesign are not mutually exclusive; they attack different layers of the pipeline.

From a systems perspective:

- Penalty-level methods change optimization geometry in coefficient space.
- Rectification changes statistical geometry in feature space before optimization.

This distinction matters for engineering reuse. If the mechanism is mainly representational, the approach can be plugged into mature sparse-learning stacks without custom solvers, while still improving support behavior in targeted dependence regimes.

6.11 WHAT THIS THEOREM DOES NOT CLAIM

To avoid overreach, the following non-claims are explicit:

- It does not prove universal improvement for arbitrary nonzero thresholds.
- It does not guarantee gains for every dataset or every correlation sign pattern.
- It does not replace empirical validation; it explains a mechanism under a tractable regime.

These caveats are consistent with prior reports that transformed and untransformed tradeoffs can vary by regime, with some datasets favoring raw discrimination while transformed models improve interpretability and attribution fidelity [5, 18].

6.12 GENERALIZATION ROADMAP BEYOND THE POC THEOREM

The PoC theorem is intentionally conservative. To move from this scoped result to broader claims, the next theoretical steps are explicit:

1. Extend analysis from zero-threshold sign mapping to learned nonzero critical ranges.
2. Characterize behavior under non-Gaussian dependence (for example heavy tails, skew, and mixture distributions).
3. Replace purely asymptotic arguments with finite-sample concentration bounds that better match practical dataset sizes.

4. Clarify edge cases where correlation contraction is weak or asymmetric, especially when negative-correlation structure dominates.

Each item maps directly to a falsifiable question in later chapters. If a proposed extension fails, the failure still sharpens the applicability boundary of the method, which is a valid and useful RQ2 outcome. If it succeeds, the dissertation moves from proof-of-concept theory toward deployable guarantees with clearer external validity.

6.13 HOW RQ2 GUIDES SUBSEQUENT CHAPTERS

This chapter provides the mechanism hypothesis that later chapters test at scale:

1. If dependence contraction is the driver, transformed models should show stronger support stability under controlled dependence stress.

2. If IC-margin effects are real, false-positive lag selection should drop in threshold-aligned regimes.

3. If claims are regime-limited, failures should cluster where assumptions break (for example nonzero-threshold mismatch or adverse correlation structure).

These predictions convert RQ2 from abstract theory into falsifiable experimental expectations.

6.14 PRACTICAL SIGNIFICANCE

For RQ2, the core result is a principled explanation for why a simple preprocessing step can improve sparse support recovery in a difficult longitudinal regime. The significance is twofold:

- **Scientific:** it links empirical improvements to established model-selection theory through IC-oriented analysis [36].

- **Engineering:** it supports a modular pipeline design where preprocessing changes correlation geometry while downstream sparse solvers remain standard and efficient.

In summary, the PoC theorem does not end the theoretical story, but it establishes a coherent bridge from binarization to IC favorability. That bridge is sufficient to motivate deeper generalization work while already explaining a large portion of observed RQ1 behavior.

CHAPTER 7

ANYTIME RULE COMPRESSION (RQ3)

Research Question 3 asks whether a sparse, rectified linear model can be compressed into a compact rule form while preserving practical discrimination quality. In this dissertation, the target is not just sparsity in coefficient space, but operational interpretability: a model that can be read, explained, and validated as an explicit decision rule under realistic constraints. The anytime rule compression framework was proposed to address this requirement by converting a fitted sparse model into an ordered sequence of candidate logic rules and allowing compression to stop as soon as performance is good enough for the application context [17, 18].

7.1 WHY RQ3 MATTERS AFTER RQ1 AND RQ2

RQ1 and RQ2 establish two foundations: first, binarization plus sparse learning can recover relevant structure more reliably in threshold-driven settings; second, the theoretical conditions for support recovery can improve when feature interactions are reframed through the binarized representation. RQ3 addresses the remaining translational gap: whether those improvements can be turned into human-auditable, low-complexity rules without giving back the predictive benefit [17, 18].

This is important for longitudinal and event-focused domains where users need a concrete trigger logic rather than a dense weight vector. In healthcare monitoring, safety surveillance, and maintenance settings, end users often need to know which few conditions jointly trigger escalation, and they need that logic in a form that is stable under repeated deployment [7, 24, 32]. In interpretability terms, this chapter focuses on the transition from post hoc explanations to inherently interpretable rule structure [15, 22].

7.2 FROM RECTIFIED SPARSE MODELS TO RULE SPACE

Given a rectified design matrix and an L1-regularized logistic model, the baseline predictor can be written as

$$\hat{p}(y_i = 1 \mid \mathbf{x}_i) = \sigma\left(\beta_0 + \sum_{j=1}^{p} \beta_j \tilde{x}_{ij}\right), \tag{1}$$

where $\tilde{x}_{ij}$ is the binarized/rectified feature and β_j is sparse due to the L1 penalty [6, 28, 29]. Let the active set be $\mathscr{A} = \{j : \beta_j \neq 0\}$. RQ3 starts from this fitted model and asks whether we can replace the full weighted sum with a smaller rule that preserves operating performance.

The compression idea is to map the active coefficients to a shared magnitude (logic polishing), keep their signs, and evaluate truncated models along an ordered path. Ordering is by absolute effect size from the trained sparse model so that early steps prioritize the most influential terms [17].

7.3 ANYTIME COMPRESSION OBJECTIVE

The objective is to optimize a practical trade-off:

$$\max_{\text{rule complexity } C} J(C) \quad \text{subject to} \quad C \leq C_{\max}, \tag{2}$$

where J is Youden's statistic ($J = \mathrm{TPR} + \mathrm{TNR} - 1$), and complexity is represented by the number of retained conditions k (and optionally the vote threshold m in an m-of-k rule) [17]. Rather than commit to a single fixed complexity in advance, the anytime procedure scans candidate compressions and can stop at any prefix as soon as the achieved J is within an application-defined tolerance of the baseline.

This aligns with a deployment reality: in some environments, a 1–2% relative loss may be acceptable if complexity falls by an order of magnitude; in others, compression is only accepted at near-zero loss. The framework supports either case by explicitly parameterizing the tolerance [17].

7.4 ALGORITHMIC PIPELINE

The compression process can be summarized in four stages [17].

7.4.1 1. Train Baseline Sparse Rectified Model

Fit L1-regularized logistic regression on the rectified matrix and compute baseline operating metrics (AUC, J, sensitivity, specificity) at the chosen operating threshold. This baseline defines the reference quality floor for compression [6, 17, 28, 29].

7.4.2 2. Logic Polishing and Ordered Prefix Construction

Restrict to active features $\mathscr{A}$ and sort them by $|\beta_j|$ descending. For prefix length k, keep the top-k signed indicators and assign common magnitude K (or equivalently use signed votes). This yields candidate rule families of increasing complexity, each nested in the next [17].

7.4.3 3. Vote Threshold Optimization

For each prefix, evaluate an m-of-k decision threshold. Intuitively, m controls how many supportive conditions are required before predicting the positive class. Sweeping m gives a local complexity-performance frontier at fixed k; sweeping k gives the global anytime frontier [17].

7.4.4 4. Adoption Rule with Relative Tolerance

Let J_{base} be the baseline score and $J(k,m)$ the compressed candidate. A candidate is accepted when

$$J(k,m) \geq (1-\varepsilon)\,J_{\text{base}}, \tag{3}$$

for configured tolerance ε. Among accepted candidates, one can choose the smallest k (maximal simplification) or the best J under a complexity cap. This makes compression policy explicit instead of ad hoc [17].

7.5 WHY THE METHOD IS "ANYTIME"

The method is anytime in the algorithmic sense: each additional step (larger k or alternative m) refines quality, but every intermediate rule is already a valid deployable model. If computation must stop early, the current incumbent can still be used. If more budget is available, scanning continues to improve operating characteristics. This property is practical for iterative model governance and constrained environments [17].

7.6 EMPIRICAL EVIDENCE TO DATE

In the prior study that introduced this framework, compression often reduced model complexity by large factors while preserving discrimination closely. Reported results showed substantial complexity drops (including cases near 50x reduction) with negligible AUC change under a practical equivalence margin, and frequently with stable or slightly improved J at the selected operating point [17]. Runtime behavior was also favorable relative to standard sparse solvers in high-dimensional rectified settings, supporting use as a post-fit compression stage [17].

Figure 5 illustrates the central behavior: *J* is tracked as a function of retained rule size *k*, creating an explicit curve from highly compressed to minimally compressed models. The chapter-level answer to RQ3 is based on this curve, not on a single arbitrarily chosen point.

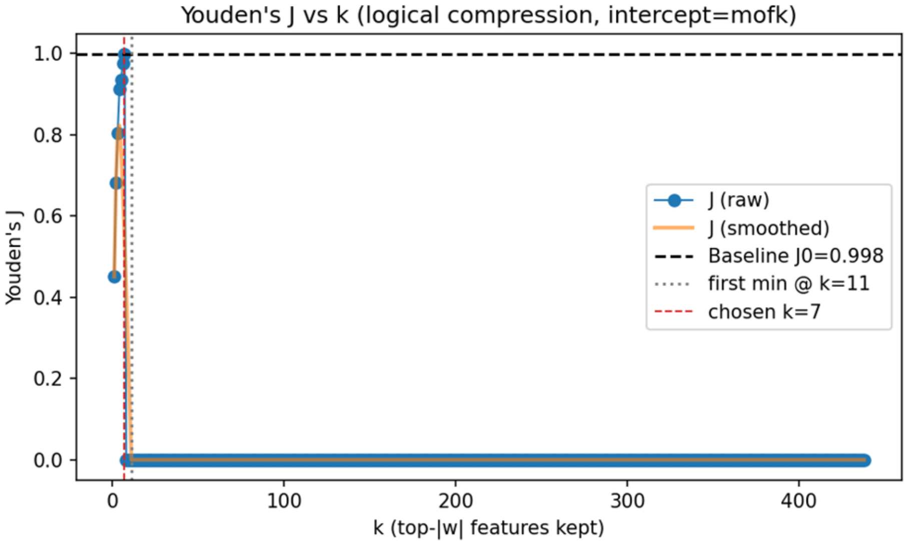

Figure 5. *J* versus retained rule size *k* for anytime compression. The curve provides a direct complexity-performance trade-off for selecting a deployable *m*-of-*k* rule.

7.7 STATISTICAL FRAMING OF “NO MATERIAL DEGRADATION”

RQ3 uses “no material degradation” in a practical, decision-oriented sense rather than a literal requirement of zero difference. Classical null-hypothesis difference tests are not sufficient to show retained utility, because failure to reject a difference is not evidence of equivalence. Accordingly, the prior work used an equivalence framing (TOST-style logic) with a predeclared margin (for example, ±0.01 AUC) to justify practical non-loss claims [17, 23, 27]. This is methodologically aligned with the chapter objective: demonstrate that simplified rules remain fit for purpose under explicit tolerance.

7.8 RELATIONSHIP TO ADJACENT INTERPRETABLE METHODS

Rule-list and scoring-system approaches provide strong interpretability baselines, but they often optimize directly in combinatorial rule space with different scalability profiles [1, 22]. The anytime compression strategy differs by leveraging a sparse convex fit first, then performing structured compression in a constrained post-processing stage. This separation is useful when the rectified feature space is large and sparse optimization is already part of the workflow.

The approach is also related to Logical Analysis of Data (LAD), particularly in its use of thresholded indicators and logical structure, but it targets a different optimization path and deployment interface [2, 3]. In the dissertation context, this can be viewed as a bridge between sparse statistical learning and logic-level decision rules tailored to threshold-driven events [18, 19].

7.9 MOTIVATION AND APPLICATION CONTEXT

The practical motivation for anytime rule compression is strongest in settings where decisions must be explained and audited quickly:

1. Clinical and public-health monitoring: produce concise trigger logic for intervention review and cross-site validation in longitudinal streams [7, 32].

2. Safety and reliability operations: support transparent condition-based escalation when false negatives are costly and operators need explicit criteria.

3. Policy-facing analytics: convert technical models into rules stakeholders can inspect, chal-

lenge, and recalibrate without retraining a complex model stack.

Across these settings, the benefit is not only readability. Compact rules also reduce implementation burden, simplify drift monitoring, and support deterministic replay during audits. The anytime mechanism then provides a tunable point on the complexity-performance frontier for each deployment constraint [15, 17].

7.10 LIMITATIONS AND SCOPE

The current evidence is strongest for the dissertation's target regime: sparse, threshold-mediated phenomena represented through rectified features. It is a proof-of-concept framework, not a universal guarantee that every sparse model admits large compression at negligible loss. Performance depends on the shape of the learned coefficient spectrum, class balance, and operating-point priorities. These boundaries are consistent with the chapter's scope and motivate the empirical analyses that follow.

7.11 INTERIM ANSWER TO RQ3

The accumulated evidence supports a positive interim answer: sparse rectified models can often be compressed into substantially smaller m-of-k rule representations with little practical degradation in discrimination, while improving interpretability and deployability. The key contribution is the anytime formulation itself: it makes compression controllable, auditable, and explicitly tied to operational tolerances rather than fixed model-size heuristics [17, 18].

CHAPTER 8

FUTURE WORK

This chapter defines the work that must be completed before the final formal dissertation defense. The goal is not to introduce a new research direction, but to close remaining validity gaps in a controlled, auditable way and convert the current proof-of-concept contributions into a defense-ready body of evidence. The plan extends the method lineage established in prior work while preserving the same three-stage architecture: critical-range rectification, sparse fitting, and anytime rule compression [17–19].

8.1 DEFENSE-READINESS OBJECTIVE

The final defense package will be considered complete when the dissertation demonstrates all three of the following:

1. **Empirical credibility:** RQ1 claims are supported by stability-aware, multi-dataset evidence rather than single-split performance snapshots.
2. **Theoretical transparency:** RQ2 assumptions, scope boundaries, and failure modes are explicit, testable, and consistent with established selection-consistency theory [5, 36].
3. **Operational interpretability:** RQ3 compression yields compact rules with documented complexity/performance tradeoffs and practical non-inferiority checks [17, 23].

These criteria align with interpretable-ML guidance that model quality must include predictive performance, descriptive fidelity, and audience relevance for high-stakes use [15, 22].

8.2 WORKSTREAM A: RQ1 EMPIRICAL STRENGTHENING

8.2.1 A1. Stability-First Evaluation Protocol

The first required expansion is to report *selection stability* and *lag fidelity* as first-class outcomes alongside AUC, Youden's J, and sparsity. This directly addresses known dependence-related fragility in sparse selection and avoids over-claiming based only on point predictive metrics [5, 36].

Planned outputs:

- Bootstrap and repeated-fold stability summaries for selected features and lags.
- False-positive and false-negative attribution summaries under controlled synthetic truth.
- Runtime and memory scaling under lag expansion.

8.2.2 A2. Baseline Expansion Under Multicollinearity

To position results more rigorously, RQ1 comparisons will include baseline families that are specifically designed for correlated predictors:

- L1 and elastic-net path baselines [6, 28, 29, 39].
- Adaptive/weighted sparse variants [20, 38].
- Grouped-penalty baselines [14, 31, 34, 35].
- Ordered lag-constrained sparse baselines for time-lag structure [30].
- Multicollinearity-focused selection comparators [10].

The objective is not to show a universal winner, but to quantify where the rectification-first pipeline helps, where it is neutral, and where it should not be preferred.

8.2.3 A3. Cross-Domain Longitudinal Validation

At least one additional longitudinal setting will be included beyond current core experiments to test transferability of the threshold-and-lag assumptions. Candidate contexts include ICS anomaly detection and biomedical/clinical trajectory settings [7, 24, 32, 33]. A legacy signal-processing dataset is retained as a secondary check for lagged-feature behavior [25].

8.3 WORKSTREAM B: RQ2 THEORETICAL COMPLETION

8.3.1 B1. Scope-Accurate Extension of the Current Derivation

The current theory is intentionally proof-of-concept. Before defense, the chapter will be finalized with a clearer bridge from the zero-threshold derivation to limited non-zero-threshold settings, while preserving explicit caveats about generality [18].

Required updates:

- Separate theorem-level claims from engineering intuition in the prose.
- Make all assumptions testable or falsifiable in synthetic studies.
- Add a concise failure-mode section for cases where dependence structure violates required conditions.

8.3.2 B2. Assumption Stress Testing

Each assumption in the RQ2 chapter will be paired with at least one targeted stress test. This is necessary because support-recovery claims are known to be sensitive to covariance geometry and signal-strength conditions [5, 36].

Planned stress tests:

- Correlation-pattern sweeps that deliberately approach known problematic IC regimes.
- Threshold-placement perturbations to measure robustness away from idealized critical ranges.
- Controlled negative-correlation regimes and complement-feature handling.

8.4 WORKSTREAM C: RQ3 ANYTIME COMPRESSION VALIDATION

8.4.1 C1. Real-Data Anytime Curves and Operating-Point Policy

RQ3 will be extended with full J-versus-k compression curves on real data and a prespecified adoption policy for practical deployment. The policy will explicitly encode acceptable complexity/performance tradeoffs instead of selecting rule size post hoc [17].

8.4.2 C2. Equivalence-Style Model Comparison

For compressed versus baseline models, evaluation will include equivalence-style testing with predefined margins rather than only “no significant difference” claims [23, 27]. This is required to support statements of practical non-loss.

8.4.3 C3. Interpretability Quality Audit

Rule models will be audited with explicit structural metrics: number of conditions, effective rule length, and decision-path simulatability. This follows interpretable-model guidance that transparent structure should be measured directly, not inferred from sparse coefficients alone [15, 22].

8.4.4 C4. Comparator Rule Learners

At least one rule-list and one logical-pattern baseline will be included for context:

- Certifiably optimal rule-list style baselines [1].
- LAD-style logical methods and implementation variants [2, 3].

This comparison is intended to clarify when anytime compression is the best practical choice versus when direct combinatorial rule learning is preferable.

8.5 CROSS-CUTTING ABLATION PROGRAM

All three research questions will use one unified ablation design to isolate causal contribution by stage:

1. Raw features + sparse learner.
2. Rectified features + sparse learner.
3. Rectified features + sparse learner + anytime compression.
4. Optional controls (for example, raw + sparse learner + compression) to separate representation from compression effects.

This program is the main mechanism for demonstrating that each stage contributes measurable value and for preventing confounding between preprocessing, fitting, and post-processing choices [17–19].

8.6 REPRODUCIBILITY AND DEFENSE ARTIFACTS

Before defense, all experiments will be migrated to a reproducible execution bundle with:

- fixed seeds, versioned splits, and immutable feature-generation settings;
- per-run manifests capturing code revision, data hash, and hyperparameter state;
- automated export of metrics, selected supports, and rule artifacts;
- one-command regeneration scripts for every figure/table used in the dissertation.

The expected outcome is a committee-auditable artifact package that can reproduce core claims without manual intervention.

8.7 RISKS AND MITIGATION PRIOR TO DEFENSE

- **Risk:** Outlier-driven critical ranges produce brittle rules. **Mitigation:** robust quantile variants and sensitivity envelopes around range boundaries [18].
- **Risk:** Correlated-feature regimes still destabilize supports. **Mitigation:** expanded correlated-baseline pack and stability reporting under repeated resampling [35, 38, 39].
- **Risk:** Compression appears accurate but changes operating behavior. **Mitigation:** operating-point diagnostics plus equivalence-margin reporting [17, 23].
- **Risk:** Overstated theoretical generality. **Mitigation:** explicit scope language and assumption-level counterexample tests [5, 36].

8.8 TIMELINE AND COMMITTEE DECISION GATES

The schedule below identifies concrete pre-defense gates. Dates are targets and may be adjusted by committee direction, but each gate has a required output and a go/no-go criterion.

8.9 PRE-DEFENSE EXIT CHECKLIST

Prior to scheduling the final defense date, the following checklist must be complete:

1. All RQ chapters include explicit claim boundaries and supporting evidence.

2. All major figures and tables regenerate from scripted workflows.

3. Citation coverage is complete for methodological claims and comparator methods.

4. At least one committee-facing artifact review confirms reproducibility.

5. Open methodological risks are documented with mitigation status.

Completion of this checklist marks transition from exploratory dissertation development to defense execution.

Table 3. Pre-defense milestone plan with required outputs and decision gates.

Milestone	Target Date	Required Output	Gate Criterion
Prospectus defense	23 Feb 2026	Approved scope and committee directives	Proceed with agreed RQ closure plan
D3 submission	06 Mar 2026	Filed prospectus documents	Administrative acceptance complete
RQ1 stability package	31 Mar 2026	Stability + ablation report draft	Metrics include stability and lag fidelity
Expanded baseline benchmarking	30 Apr 2026	Multicollinearity baseline comparison	No missing baseline family category
RQ2 assumption stress tests	29 May 2026	Theory-scope and failure-mode supplement	Assumptions mapped to explicit tests
RQ3 real-data compression package	19 Jun 2026	Anytime curves + equivalence analysis	Compression policy fully prespecified
Full draft to committee	03 Jul 2026	Complete dissertation manuscript	All chapters internally consistent
Mock defense and revision closeout	10 Jul 2026	Final slide deck and resolved action log	No critical unresolved findings
Formal dissertation defense	17 Jul 2026	Defended dissertation	Committee pass with required revisions
Final submission and graduation	28 Aug 2026	Post-defense corrections and deposit	Degree conferral

CHAPTER 9

CONCLUSION

This dissertation examined longitudinal feature learning under three simultaneous constraints: dependence induced by lag expansion, the need for reliable feature-lag attribution, and the requirement for compact, human-auditable model outputs. The central claim is that a representation-first pipeline, followed by sparse fitting and anytime rule compression, provides a practical middle ground between unstable raw sparse selection and computationally heavy direct rule search [17–19].

9.1 RESEARCH PROGRAM SUMMARY

The work was organized around three research questions:

- **RQ1:** whether critical-range rectification improves sparse attribution reliability in longitudinal settings;
- **RQ2:** why rectification helps under a defensible theoretical mechanism and where that mechanism is scoped;
- **RQ3:** whether sparse rectified models can be compressed into compact logical rules without unacceptable operational loss.

Together, these questions move from empirical behavior (RQ1), to theoretical plausibility (RQ2), to deployment usability (RQ3). This sequencing is intentional: interpretability claims are weak if support recovery is unstable, and practical rule compression is less meaningful without a credible upstream selection mechanism [5, 15, 22, 36].

9.2 CONCLUSIONS BY RESEARCH QUESTION

9.2.1 RQ1 Conclusion: Rectification Can Improve Attribution Reliability in Target Regimes

The accumulated evidence supports a conditional positive conclusion for RQ1. In threshold-and-lag aligned settings, critical-range rectification tends to improve support concentration, lag localization, and sparse-model usability under multicollinearity stress [18, 19]. This finding is consistent with known dependence-related limits of penalty-only sparse selection, where prediction can remain acceptable while support identification degrades [5, 36].

The dissertation does not claim universal dominance over all baseline families. Instead, it establishes that representation-level intervention is a viable and often effective alternative to penalty redesign alone, especially when events are naturally threshold-mediated.

9.2.2 RQ2 Conclusion: A Scoped IC-Based Mechanism is Theoretically Defensible

For RQ2, this work provides a proof-of-concept theoretical argument: in a tractable regime (zero-threshold sign binarization with explicit assumptions), dependence contraction can improve the likelihood of satisfying IC-related sparse-recovery conditions [18, 36]. The theoretical contribution is therefore mechanistic rather than universal.

This distinction is central to rigor. The chapter argues for a plausible structural pathway from binarization to improved sparse-support behavior, while explicitly separating theorem-scoped claims from broader engineering intuition. The resulting conclusion is that rectification is not just a heuristic preprocessing trick; it has a coherent connection to established sparse-selection theory under defined conditions.

9.2.3 RQ3 Conclusion: Anytime Compression Makes Interpretability Operational

For RQ3, this work concludes that sparse rectified models can often be compressed into substantially smaller rule representations while preserving practical discrimination quality under predefined tolerance policies [17, 18]. This directly addresses deployment requirements in settings where explanation, auditability, and compact decision logic are non-negotiable [15, 22].

The key advance is the anytime framing itself: rule complexity can be reduced incrementally with explicit stop/adoption criteria, making the interpretability-performance tradeoff controllable rather than ad hoc. Equivalence-oriented evaluation framing further supports practical non-loss claims when simplification is the objective [23, 27].

9.3 PRIMARY DISSERTATION CONTRIBUTIONS

The dissertation contributes the following integrated advances:

1. **A rectification-first longitudinal feature-learning pipeline** that preserves lag structure while mapping signals into critical-range indicators with clear operational semantics [18, 19].

2. **A scoped theoretical bridge to sparse-recovery conditions** showing how representation-level dependence contraction can increase IC favorability in a proof-of-concept regime [18, 36].

3. **An anytime rule-compression stage** that converts sparse coefficients into compact *m*-of-*k*-style logic with tunable performance-complexity tradeoffs [17].

4. **A unified evaluation perspective** that balances discrimination, attribution stability, interpretability complexity, and computational efficiency rather than optimizing a single metric.
5. **A reproducible implementation pathway** for rectification, sparse fitting, and compression, enabling repeatable comparisons and practical extension in follow-on work.

9.4 POSITIONING RELATIVE TO PRIOR WORK

This work extends, rather than replaces, prior sparse and interpretable modeling literature. Relative to classical L1 and elastic-net families [6, 28, 29, 38, 39], the main difference is intervention point: representation is modified before sparse optimization. Relative to grouped and ordered penalties [14, 30, 31, 34, 35], this work emphasizes threshold logic and lag-attribution clarity in a longitudinal context. Relative to direct rule-learning approaches [1–3], it uses sparse convex fitting as a scalable front end and performs structured rule simplification afterward.

Accordingly, the work is best interpreted as a bridge architecture between sparse statistical learning and rule-level interpretability, not as a claim that one method family should replace all others.

9.5 SCOPE BOUNDARIES AND LIMITATIONS

Several limitations remain explicit:

- Benefits are strongest in regimes with threshold-mediated event structure and may be weaker when that structure is absent.
- The current theory is proof-of-concept and does not provide universal guarantees across arbitrary threshold schemes or dependence geometries.

- Compression quality depends on operating-point policy, class balance, and the shape of the learned sparse support.
- Interpretability gains can trade against raw discrimination in some datasets.

These boundaries are consistent with broader evidence that no sparse-selection or interpretable-modeling method is uniformly optimal across all dependence structures and objectives [5, 15].

9.6 PRACTICAL IMPLICATIONS

This work's practical implication is straightforward: when longitudinal decisions require both predictive utility and audit-ready rationale, it can be more effective to restructure features around event-relevant critical ranges first, then apply mature sparse optimization, and finally compress to explicit rule logic. This design is particularly relevant in domains such as ICS monitoring and clinical trajectory analysis, where lag effects, threshold triggers, and traceable decision criteria are operationally important [7, 24, 32].

9.7 FINAL STATEMENT

This work concludes that rectification-first sparse longitudinal learning with anytime rule compression is a credible and useful framework for interpretable event modeling under dependence. Its value lies in integration: empirical support behavior (RQ1), scoped mechanistic theory (RQ2), and operational rule simplification (RQ3) are addressed as a single pipeline rather than isolated techniques [17–19]. Within its stated scope, this provides a defensible contribution to interpretable machine learning for longitudinal data.

REFERENCES

[1] E. Angelino, N. Larus-Stone, D. Alabi, M. Seltzer, and C. Rudin, “Learning certifiably optimal rule lists for categorical data,” *Journal of Machine Learning Research*, 2018.

[2] E. Boros, P. L. Hammer, T. Ibaraki, and A. Kogan, “Logical analysis of numerical data,” *Mathematical Programming*, 1997.

[3] E. Boros, P. L. Hammer, T. Ibaraki, A. Kogan, E. Mayoraz, and I. Muchnik, “An implementation of logical analysis of data,” *IEEE Transactions on Knowledge and Data Engineering*, 2000.

[4] W. Feller, *An Introduction to Probability Theory and Its Applications*. Wiley, 1991.

[5] L. Freijeiro-González, M. Febrero-Bande, and W. González-Manteiga, “A critical review of lasso and its derivatives for variable selection under dependence among covariates,” *International Statistical Review*, vol. 90, pp. 118–145, 2022.

[6] J. Friedman, R. Tibshirani, and T. Hastie, “Regularization paths for generalized linear models via coordinate descent,” *Journal of Statistical Software*, vol. 33, pp. 1–22, 2010. DOI: 10.18637/jss.v033.i01

[7] C. Hiploylee, P. Dufort, H. Davis, R. Wennberg, M. Tartaglia, D. Mikulis, and C. Tator, “Longitudinal study of postconcussion syndrome: Not everyone recovers,” *Journal of Neurotrauma*, vol. 34, pp. 1511–1523, 2017.

[8] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 2012.

[9] T. Inoue, H. Qiu, R. Ueji, and Y. Kimura, “Ductile-to-brittle transition and brittle fracture stress of ultrafine-grained low-carbon steel,” *Materials*, vol. 14, p. 1634, 2021.

[10] A. Katrutsa and V. Strijov, “Comprehensive study of feature selection methods to solve multicollinearity problem according to evaluation criteria,” *Expert Systems with Applications*, vol. 76, pp. 1–11, 2017.

[11] Y. Kim, J. Kim, and Y. Kim, “Blockwise sparse regression,” *Statistica Sinica*, pp. 375–390, 2006.

[12] M. KP and P. Thiyagarajan, “Feature selection using efficient fusion of fisher score and greedy searching for alzheimer’s classification,” *Journal of King Saud University-Computer and Information Sciences*, vol. 34, pp. 4993–5006, 2022.

[13] L. Ladha and T. Deepa, “Feature selection methods and algorithms,” *International journal on computer science and engineering*, vol. 3, pp. 1787–1797, 2011.

[14] L. Meier, S. VanDeGaer, and P. Bühlmann, “The group lasso for logistic regression,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 70, pp. 53–71, 2006.

[15] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, “Definitions, methods, and applications in interpretable machine learning,” *Proceedings of the National Academy of Sciences*, 2019.

[16] A. Y. Ng, “Feature selection, l1 vs. l2 regularization, and rotational invariance,” in *Proc. ICML*, 2004.

[17] J. Orender, J. Sun, and M. Zubair, “Anytime rule compression and rectified logistic modeling for longitudinal signals,” in *1st International Conference on Big Data Analytics and Applications (BDAA)*, 2025.

[18] J. Orender, J. Sun, and M. Zubair, “Efficient longitudinal feature selection via binarized transformation: Theory and case studies,” in *IEEE Big Data*, 2025.

[19] J. Orender, M. Zubair, and J. Sun, “Lasso logic engine: Harnessing the logic parsing capabilities of the lasso algorithm for longitudinal feature learning,” in *2022 IEEE International Conference on Big Data (Big Data)*, IEEE, 2022, pp. 309–318.

[20] M. Rejchel and M. Bogdan, “Rank-based lasso: Efficient methods for high-dimensional robust model selection,” *Journal of Machine Learning Research*, 2020.

[21] M. Roozbeh, S. Babaie-Kafaki, and A. Sadigh, “A heuristic approach to combat multicollinearity in least trimmed squares regression analysis,” *Applied Mathematical Modelling*, vol. 57, pp. 105–120, 2018.

[22] C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” *Nature Machine Intelligence*, 2019.

[23] D. J. Schuirmann, “A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability,” *Journal of Pharmacokinetics and Biopharmaceutics*, 1987.

[24] H. Shin, W. Lee, J. Yun, and H. Kim, “Hai 1.0: Hil-based augmented ics security dataset,” in *13th USENIX Workshop on Cyber Security Experimentation and Test*, Santa Clara, CA, USA: CSET, 2020.

[25] V. Sigillito, S. Wing, L. Hutton, and K. Baker, “Classification of radar returns from the ionosphere using neural networks,” *Johns Hopkins APL Technical Digest*, vol. 10, pp. 262–266, 1989.

[26] J. Sorvali, J. Kaseva, and P. Peltonen-Sainio, “Farmer views on climate change—a longitudinal study of threats, opportunities and action,” *Climatic Change*, vol. 164, pp. 1–19, 2021.

[27] Student, “The probable error of a mean,” *Biometrika*, 1908.

[28] J. Tay, B. Narasimhan, and T. Hastie, “Elastic net regularization paths for all generalized linear models,” *Journal of Statistical Software*, vol. 106, pp. 1–31, 2023. DOI: `10.18637/jss.v106.i01`

[29] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, pp. 267–288, 1996.

[30] R. Tibshirani and X. Suo, “An ordered lasso and sparse time-lagged regression,” *Technometrics*, 2016.

[31] B. Turlach, W. Venables, and S. Wright, “Simultaneous variable selection,” *Technometrics*, vol. 47, pp. 349–363, 2005.

[32] C. Wong, A. Caspi, B. Williams, I. Craig, R. Houts, A. Ambler, T. Moffitt, and J. Mill, “A longitudinal study of epigenetic variation in twins,” *Epigenetics*, vol. 5, pp. 516–526, 2010.

[33] L. Yang and T. Wu, “Model-based clustering of high-dimensional longitudinal data via regularization,” *Biometrics*, vol. 79, pp. 761–774, 2023.

[34] Y. Yang and H. Zou, “A fast unified algorithm for solving group-lasso penalize learning problems,” *Statistics and Computing*, vol. 25, pp. 1129–1141, 2015.

[35] M. Yuan and Y. Lin, “Model selection and estimation in regression with grouped variables,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, pp. 49–67, 2006.

[36] P. Zhao and B. Yu, "On model selection consistency of lasso," *Journal of Machine Learning Research*, 2006.

[37] H. Zhu, W. Xu, Z. Zhang, Y. Xu, and J. Fan, "A variable selection approach for highly correlated predictors in high-dimensional genomic data," *Bioinformatics*, 2021.

[38] H. Zou, "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, 2006.

[39] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B*, 2005.

APPENDIX A

TOY EXAMPLE

A.1 DATA

The toy example data demonstrates the behavior described in the paper. A matrix with features 'A' through 'J' shows that even with only 10 features, the coefficients become muddled. Columns A and D were designed to produce the result vector, while the others were largely random, with top elements tweaked to ensure they are **not** optimal for producing the result vector.

$$\mathbf{X} = \left[\begin{array}{cccccccccc}
\text{A} & \text{B} & \text{C} & \text{D} & \text{E} & \text{F} & \text{G} & \text{H} & \text{I} & \text{J} \\
0.49 & 0.48 & 0.42 & 0.13 & 0.45 & 0.48 & 0.42 & 0.48 & 0.43 & 0.46 \\
0.12 & 0.52 & 0.46 & 0.47 & 0.42 & 0.47 & 0.49 & 0.47 & 0.51 & 0.48 \\
0.23 & 0.11 & 0.38 & 0.57 & 0.99 & 0.51 & 0.41 & 0.71 & 0.29 & 0.56 \\
\mathbf{0.48} & 0.40 & 0.59 & \mathbf{0.46} & 0.41 & 0.43 & 0.40 & 0.41 & 0.60 & 0.52 \\
0.83 & 0.45 & 0.48 & 0.54 & 0.44 & 0.44 & 0.23 & 0.41 & 0.46 & 0.49 \\
\mathbf{0.55} & 0.52 & 0.41 & \mathbf{0.58} & 0.40 & 0.60 & 0.51 & 0.42 & 0.42 & 0.45 \\
0.53 & 0.02 & 0.55 & 0.16 & 0.38 & 0.06 & 0.50 & 0.55 & 0.71 & 0.51 \\
\mathbf{0.54} & 0.60 & 0.56 & \mathbf{0.47} & 0.49 & 0.52 & 0.52 & 0.56 & 0.54 & 0.58 \\
0.33 & 0.22 & 0.59 & 0.49 & 0.47 & 0.49 & 0.59 & 0.47 & 0.91 & 0.57 \\
0.34 & 0.49 & 0.66 & 0.51 & 0.48 & 0.79 & 0.84 & 0.53 & 0.73 & 0.67
\end{array}\right]$$

A.2 TRANSFORMATION

After transformation, the matrix is encoded with ones and negative ones. The encoding rule was simple: values between 0.4 and 0.6 (inclusive) became ones; others became negative ones. Details on defining this *critical range* are in the other sections. In summary, if both features A and D were between 0.4 and 0.6 simultaneously, an event occurred.

$$\mathbf{X} = \left[\begin{array}{cccccccccc}
\mathrm{A} & \mathrm{B} & \mathrm{C} & \mathrm{D} & \mathrm{E} & \mathrm{F} & \mathrm{G} & \mathrm{H} & \mathrm{I} & \mathrm{J} \\
1 & 1 & 1 & -1 & 1 & 1 & 1 & 1 & 1 & 1 \\
-1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
-1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 & -1 & 1 \\
\mathbf{1} & 1 & 1 & \mathbf{1} & 1 & 1 & 1 & 1 & 1 & 1 \\
-1 & 1 & 1 & 1 & 1 & 1 & -1 & 1 & 1 & 1 \\
\mathbf{1} & 1 & 1 & \mathbf{1} & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & -1 & 1 & -1 & -1 & -1 & 1 & 1 & -1 & 1 \\
\mathbf{1} & 1 & 1 & \mathbf{1} & 1 & 1 & 1 & 1 & 1 & 1 \\
-1 & -1 & 1 & 1 & 1 & 1 & 1 & 1 & -1 & 1 \\
-1 & 1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 & -1
\end{array}\right]$$

Here is the boolean response vector:

$$y^T = \begin{bmatrix} F & F & F & \mathbf{T} & F & \mathbf{T} & F & \mathbf{T} & F & F \end{bmatrix}$$

It can be shown that the only possible logical AND combination of these features which would produce the response vector shown would be the 1st and 4th columns.

APPENDIX B

LEMMAS

B.1 LEMMA 1

The relationship between the pre-binarization correlation (ρ), and the post binarization correlation ($\tilde{\rho}$) for **standard normal** random variables is:

$$\tilde{\rho} = \frac{2}{\pi} \arcsin(\rho)$$

Step 1. *Basic definitions.*

Let X and Y be **standard normal** random variables with:

- **Means:** $E[X] = E[Y] = 0$ (centered)
- **Variances:** $\mathrm{Var}(X) = \mathrm{Var}(Y) = 1$ (standardized)
- **Correlation:** $\mathrm{Corr}(X,Y) = \rho$, where $-1 \leq \rho \leq 1$

Step 2. *Define the Binarized Variables.*

We binarize X and Y using the threshold zero:

$$\tilde{X} = \begin{cases} 1 & \text{if } X > 0 \\ 0 & \text{if } X \leq 0 \end{cases} \quad \text{and} \quad \tilde{Y} = \begin{cases} 1 & \text{if } Y > 0 \\ 0 & \text{if } Y \leq 0 \end{cases}$$

This $0/1$ encoding is affine-equivalent to the $\{\pm 1\}$ sign encoding used in Chapter 06; after centering, the correlation-ordering arguments are unchanged up to constant scaling.

Step 3. *Calculate Means and Variances of Binarized Variables.*

Since X and Y are symmetric about zero:

- Mean of $\tilde{X} : E[\tilde{X}] = P(X > 0) = 0.5$
- Variance of $\tilde{X} : \text{Var}(\tilde{X}) = E[\tilde{X}^2] - (E[\tilde{X}])^2 = 0.5 - (0.5)^2 = 0.25$
- Similarly for $\tilde{Y} : E[\tilde{Y}] = 0.5, \quad \text{Var}(\tilde{Y}) = 0.25$

Step 4. *Express the Correlation Between Binarized Variables.*

The correlation $\tilde{\rho}$ between $\tilde{X}$ and $\tilde{Y}$ is:

$$\tilde{\rho} = \frac{E[\tilde{X}\tilde{Y}] - E[\tilde{X}]E[\tilde{Y}]}{\sqrt{\text{Var}(\tilde{X})\text{Var}(\tilde{Y})}} = 4\left(P(\tilde{X} = 1, \tilde{Y} = 1) - 0.25\right)$$

Step 5. *Define the probabilities.*

Since X and Y are assumed to be jointly normally distributed, $P(\tilde{X} = 1, \tilde{Y} = 1)$ can be expressed using the **bivariate standard normal cumulative distribution function (CDF)**.

$$P(\tilde{X} = 1, \tilde{Y} = 1) = P(X > 0, Y > 0)$$

Step 6. *Use the Bivariate Normal CDF.*

For standard normal variables X and Y with correlation ρ an expression of the *Gaussian Copula Formula* holds [4]:

$$P(X > 0, Y > 0) = \frac{1}{4} + \frac{\arcsin(\rho)}{2\pi}$$

Explanation:

- The joint probability $P(X > 0, Y > 0)$ corresponds to the *orthant probability* in the positive quadrant.

- The term $\arcsin(\rho)$ arises from integrating the bivariate normal density over this quadrant.

Step 7. *Substitute Back to Find $\tilde{\rho}$.*

$$\tilde{\rho} = 4\left(\frac{1}{4} + \frac{\arcsin(\rho)}{2\pi} - 0.25\right) = \frac{2}{\pi}\arcsin(\rho)$$

B.2 LEMMA 2

$$|\tilde{\rho}| \leq |\rho|$$

Interpretation: Using the result of Lemma 1, the ratio $\frac{\tilde{\rho}}{\rho} = \frac{2}{\pi\rho}\arcsin(\rho)$ is always less than one for ρ in the interval $(-1,1)$ except at $\rho = \pm 1$ where the ratio is equal to one.

Step 1

Properties of the Arcsine (sin^{-1}) Function:

- The **arcsine function** $\arcsin(x)$ is defined for $x \in [-1,1]$.
- The range of $\arcsin(x)$ is $[-\frac{\pi}{2}, \frac{\pi}{2}]$.
- The function $\arcsin(x)$ is **increasing** on $[-1,1]$.

Step 2

We can consider ρ in one of three intervals:

1. $\rho = 0$
2. $0 < \rho < 1$
3. $-1 < \rho < 0$

Step 3 (Case 1: $\rho = 0$)

- $\arcsin(0) = 0$
- The ratio becomes indeterminate ($\frac{0}{0}$), but we can consider the limit as $\rho \to 0$.
- As $\rho \to 0$, $\arcsin(\rho) \approx \rho$ (using the first-order Taylor expansion).
- Therefore, $\frac{\tilde{\rho}}{\rho} \approx \frac{2}{\pi\rho} \cdot \rho = \frac{2}{\pi} \approx 0.6366$.

Step 4 (Case 2: $0 < \rho < 1$**)**

Both ρ and $\arcsin(\rho)$ are positive.

Define the function:

$$f(\rho) = \frac{2}{\pi\rho} \arcsin(\rho)$$

We need to show that $f(\rho) < 1$ for $0 < \rho < 1$.

For $0 < \rho < 1$: $\arcsin(\rho) < \frac{\pi}{2}\rho$

This inequality holds because of the well known result that the function $\arcsin(\rho)$ is **strictly convex** on $(0,1)$, and its graph lies below the straight line $y = \frac{\pi}{2}\rho$ [18].

Therefore: $f(\rho) < 1$ for all $0 < \rho < 1$.

Step 5 (Case 3: $-1 < \rho < 0$**)**

- When ρ is negative $\arcsin(\rho)$ is negative, also implying that $\tilde{\rho}$ is negative.
- The ratio $\frac{\tilde{\rho}}{\rho}$ is therefore positive because a negative divided by negative is positive.
- Since $\arcsin(\rho)$ is negative and increasing from $-\frac{\pi}{2}$ to 0 as ρ increases from -1 to 0, we can use logic similar to Case 2.
- Compute the Limit as $\rho \to -1^{+}$.
- $\arcsin(-1) = -\frac{\pi}{2}$
- $f(-1) = \frac{2}{\pi \times (-1)} \times \left(-\frac{\pi}{2}\right) = 1$

- For $-1 < \rho < 0$: $\arcsin(\rho) > \frac{\pi}{2}\rho$

(using the same logic as in case 2, except that the signs are opposite)

$f(\rho) < 1$ for all $-1 < \rho < 0$.

Step 6

- The ratio $\frac{\tilde{\rho}}{\rho} = \frac{2}{\pi\rho}\arcsin(\rho)$ is **less than one** for all ρ in $(-1, 1)$.
- The ratio equals **one only when** $\rho = 1$ or $\rho = -1$.
- Therefore, **the ratio is always less than or equal to one**.
- Furthermore, since $\tilde{\rho}$ will always have the same sign as ρ, the magnitude of $\tilde{\rho}$ will always be less than the magnitude of ρ

$$|\tilde{\rho}| \leq |\rho|$$

Implications

- Under this result, binarization reduces or at worst preserves correlation - never inflates it beyond its original absolute value. That is, for any pair of binarized features, the magnitude of correlation coefficient $\tilde{\rho}$ is strictly less than the original ρ, except at ±1 where they are equal.

- This result effectively keeps all off-diagonal correlations from saturating to exactly ±1 when no column is an exact multiple of another.

- If all off-diagonal entries of a correlation matrix are strictly bounded away from ±1, then the matrix's smallest eigenvalue $\lambda_{min}(R)$ is always positive [8].

- Therefore, R^{-1} is well-defined and has a finite spectral norm bound given by:

  $\|R^{-1}\|_2 \leq \frac{1}{\lambda_{min}(R)}$.

- Hence, the boundedness of R^{-1} is implied when $|\tilde{\rho}| \leq |\rho|$ and $\tilde{\rho} \neq \pm 1$ and the entire correlation matrix must be invertible, with binarization universally enhancing its conditioning for more stable inversion in practice.

- Thus, a matrix-level corollary can be extended from the pairwise-level bound shown above. Both statements reinforce the same theme: binarization controls correlations and preserves invertibility in a way beneficial to sparse regression methods like LASSO.

B.3 LEMMA 3

$$\left|\mathrm{Cov}(\tilde{x}_{S^c i}, \tilde{x}_{S j})\right| \leq \left|\mathrm{Cov}(x_{S^c i}, x_{S j})\right|$$

Interpretation: Binarization reduces the covariance between features for any irrelevant feature $x_{S^c i} \in X_{S^c}$ and relevant feature $x_{S j} \in X_S$.

Review of Assumptions:

1. **Joint Normality:** The continuous features $x_{S^c i}$ and $x_{S j}$ are jointly normally distributed with zero means and unit variances:

$$x_{S^c i}, x_{S j} \sim \mathcal{N}(0,1)$$

   and correlation $\rho_{ij} = \mathrm{Corr}(x_{S^c i}, x_{S j})$.

2. **Binarization at Zero Threshold** - The binarized variables are defined as:

$$\tilde{X}_k = \begin{cases} 1 & \text{if } X_k > 0 \\ 0 & \text{if } X_k \leq 0 \end{cases}, \quad k = i, j$$

Step 1

The covariance between $x_{S^c i}$ and $x_{S j}$ is:

$$\mathrm{Cov}(x_{S^c i}, x_{S j}) = \rho_{ij} \cdot \sigma_{x_{S^c i}} \sigma_{x_{S j}} = \rho_{ij} \cdot (1)(1) = \rho_{ij}$$

Step 2

The covariance between $\tilde{x}_{S^c i}$ and $\tilde{x}_{S j}$ is:

$$\mathrm{Cov}(\tilde{x}_{S^c i}, \tilde{x}_{S j}) = \mathbb{E}[\tilde{x}_{S^c i} \tilde{x}_{S j}] - \mathbb{E}[\tilde{x}_{S^c i}] \mathbb{E}[\tilde{x}_{S j}]$$

Since $x_{S^c i}$ and $x_{S j}$ are symmetric about zero:

$$\mathbb{E}[\tilde{x}_{S^c i}] = P(x_{S^c i} > 0) = \frac{1}{2}, \quad \mathbb{E}[\tilde{x}_{S j}] = \frac{1}{2}$$

Compute the joint probability:

$$\mathbb{E}[\tilde{x}_{S^c i} \tilde{x}_{S j}] = P(x_{S^c i} > 0, x_{S j} > 0)$$

For jointly standard normal variables, this joint probability is given by [4]:

$$P(x_{S^c i} > 0, x_{S j} > 0) = \frac{1}{4} + \frac{\arcsin(\rho_{ij})}{2\pi}$$

Therefore, the covariance is:

$$\mathrm{Cov}(\tilde{x}_{S^c i}, \tilde{x}_{S j}) = \left(\frac{1}{4} + \frac{\arcsin(\rho_{ij})}{2\pi}\right) - \left(\frac{1}{2} \cdot \frac{1}{2}\right) = \frac{\arcsin(\rho_{ij})}{2\pi}$$

Step 2

Comparing Covariances Before and After Binarization

We need to show that:

$$\left|\frac{\arcsin(\rho_{ij})}{2\pi}\right| \leq |\rho_{ij}|$$

Case 1: When $\rho_{ij} \geq 0$

The function $f(\rho) = \frac{\arcsin(\rho)}{2\pi}$ is increasing in $[0,1]$. Since $\arcsin(\rho) \leq \frac{\pi}{2}$, we have:

$$0 \leq \frac{\arcsin(\rho_{ij})}{2\pi} \leq \frac{1}{4}$$

Also, $\rho_{ij} \leq 1$, so:

$$0 \leq \frac{\arcsin(\rho_{ij})}{2\pi} \leq \rho_{ij}$$

because $\arcsin(\rho_{ij}) \leq \frac{\pi}{2}\rho_{ij}$ in $[0,1]$.

Step 3

Case 2: When $\rho_{ij} \leq 0$

The function $f(\rho) = \frac{\arcsin(\rho)}{2\pi}$ is increasing in $[-1,0]$. Since $\arcsin(\rho_{ij}) \geq -\frac{\pi}{2}$, we have:

$$-\frac{1}{4} \leq \frac{\arcsin(\rho_{ij})}{2\pi} \leq 0$$

Also, $\rho_{ij} \geq -1$, so:

$$\rho_{ij} \leq \frac{\arcsin(\rho_{ij})}{2\pi} \leq 0$$

because $\arcsin(\rho_{ij}) \geq \frac{\pi}{2}\rho_{ij}$ in $[-1,0]$.

Therefore, in both cases:

$$\left|\mathrm{Cov}(\tilde{x}_{S^c i}, \tilde{x}_{Sj})\right| = \left|\frac{\arcsin(\rho_{ij})}{2\pi}\right| \leq |\rho_{ij}| = \left|\mathrm{Cov}(x_{S^c i}, x_{Sj})\right|$$

Conclusion

The magnitude of the covariance between $\tilde{x}_{S^c i}$ and $\tilde{x}_{Sj}$ is less than or equal to the magnitude of the covariance between $x_{S^c i}$ and x_{Sj}. Therefore, binarization reduces the covariances between irrelevant and relevant features.

B.4 LEMMA 4

When $\tilde{\rho}$ and $\rho > 0$,

$$\|\tilde{G}^{-1}\|_\infty \leq \|G^{-1}\|_\infty$$

Interpretation: The Infinity norm of the binarized matrix is less than or equal to the Infinity norm of the original continuous matrix.

Definitions:

- $n > 0$: Positive scalar (e.g., sample size).
- $s \geq 2$: Dimension of the square matrices ($s \times s$).
- $\mathbf{I}_s$: Identity matrix of size s.
- $\mathbf{J}_s$: All-ones matrix of size s.
- $0 < \tilde{\rho} < \rho < 1$: Constants with $\tilde{\rho}$ always less than ρ.

Step 1

Original Gram Matrix:

$$G = n\left((1-\rho)\mathbf{I}_s + \rho\mathbf{J}_s\right)$$

Binarized Gram Matrix:

$$\tilde{G} = n\left((1-\tilde{\rho})\mathbf{I}_s + \tilde{\rho}\mathbf{J}_s\right)$$

Inverse of G and $\tilde{G}$:

First, rewrite G as

$$G = n(1-\rho)\left(\mathbf{I}_s + \frac{\rho}{1-\rho}\mathbf{J}_s\right)$$

Let $\alpha = \dfrac{\rho}{1-\rho}$.

Compute $(\mathbf{I}_s + \alpha\mathbf{J}_s)^{-1}$:

- Observation: $\mathbf{J}_s$ is rank-1.

- Sherman-Morrison-Woodbury formula for the inverse:

$$(\mathbf{A} + \mathbf{U}\mathbf{C}\mathbf{V}^T)^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}(\mathbf{C}^{-1} + \mathbf{V}^T\mathbf{A}^{-1}\mathbf{U})^{-1}\mathbf{V}^T\mathbf{A}^{-1}$$

- $\mathbf{A} = \mathbf{I}_s \Rightarrow \mathbf{A}^{-1} = \mathbf{I}_s$

- $\mathbf{U} = \mathbf{V} = \mathbf{1}_s$

- $\mathbf{C} = \alpha$

- $\mathbf{C}^{-1} = \frac{1}{\alpha}$ (the inverse of a scalar being its reciprocal)

- $\mathbf{V}^T\mathbf{A}^{-1}\mathbf{U} = \mathbf{1}_s^T\mathbf{1}_s = s$

For $\mathbf{I}_s + \alpha\mathbf{J}_s$:

$$(\mathbf{I}_s + \alpha\mathbf{J}_s)^{-1} = \mathbf{I}_s - \frac{\alpha}{1+\alpha s}\mathbf{J}_s$$

Therefore,

$$G^{-1} = \frac{1}{n(1-\rho)} \left(\mathbf{I}_s - \frac{\rho}{1+\rho(s-1)} \mathbf{J}_s \right)$$

Similarly, for $\tilde{G}$:

$$\tilde{G}^{-1} = \frac{1}{n(1-\tilde{\rho})} \left(\mathbf{I}_s - \frac{\tilde{\rho}}{1+\tilde{\rho}(s-1)} \mathbf{J}_s \right)$$

Step 2

Definition of Infinity Norm for Matrices:

$$\|A\|_\infty = \max_{1 \le i \le s} \sum_{j=1}^{s} |a_{ij}|$$

Compute Entries of G^{-1}:

- Diagonal Entries: $(G^{-1})_{ii} = \frac{1}{n(1-\rho)} (1-\gamma)$

- Off-Diagonal Entries: $(G^{-1})_{ij} = -\frac{\gamma}{n(1-\rho)} \quad \text{for } i \neq j$

Where: $\gamma = \frac{\rho}{1+\rho(s-1)}$

Compute the Row Sum for G^{-1}:

Row Sum:

$$\text{RowSum}_{G^{-1}} = \frac{1}{n(1-\rho)} [(1-\gamma) + (s-1)\gamma]$$

Simplify the Row Sum:

$$\text{RowSum}_{G^{-1}} = \frac{1}{n(1-\rho)} [1 - \gamma + \gamma(s-1)] = \frac{1}{n(1-\rho)} [1 + \gamma(s-2)]$$

Where: $\tilde{\gamma} = \frac{\tilde{\rho}}{1+\tilde{\rho}(s-1)}$

Similarly:

$$\text{RowSum}_{\tilde{G}^{-1}} = \frac{1}{n(1-\tilde{\rho})}\left[1+\tilde{\gamma}(s-2)\right]$$

Step 3

Establishing the Relationship between γ and $\tilde{\gamma}$

Define function $f(x) = \frac{x}{1+x(s-1)}$

Due to the monotinicity of f(x), when $x_1 < x_2$, $f(x_1) < f(x_2)$

It follows that:

- $f(x)$ is *increasing* wrt x on $\left(\frac{-1}{s-1}, 1\right)$.
- $|\tilde{\rho}| \leq |\rho|$ (Lemma 2), and both are bounded by the interval $(-1,1)$
- Since $\tilde{\rho} \geq \rho$ when $\tilde{\rho}$ and ρ are less than zero (meaning that ρ is *more* negative), the monotincity of f(x) implies that $|f(\tilde{\rho})| \leq |f(\rho)|$
- Therefore, we have $|\tilde{\gamma}| = |f(\tilde{\rho})| \leq |f(\rho)| = |\gamma|$, with a discontinuity at $x = \frac{-1}{s-1}$.
- Furthermore, $\tilde{\rho}, \rho$, and hence γ, $\tilde{\gamma}$ will all be of the same sign in the interval $\left[\frac{-1}{s-1}, 1\right)$ for $\tilde{\rho}$ and ρ, and γ will be positive otherwise.

Step 4

Comparing Row Sums:

Since $|\gamma| \geq |\tilde{\gamma}|$ and $s-2 \geq 0$ (because $s \geq 2$),

$$\gamma(s-2) \geq \tilde{\gamma}(s-2) \text{ when } \gamma > 0$$

Therefore,

$$1+\gamma(s-2) \geq 1+\tilde{\gamma}(s-2) \text{ when } \gamma > 0$$

Comparing Denominators:

Since $|\tilde{\rho}| \leq |\rho|$,

- $1-\tilde{\rho} \geq 1-\rho \Rightarrow \frac{1}{n(1-\tilde{\rho})} \leq \frac{1}{n(1-\rho)}$ when $\tilde{\rho}, \rho > 0$

- $1-\tilde{\rho} \leq 1-\rho \Rightarrow \frac{1}{n(1-\tilde{\rho})} \geq \frac{1}{n(1-\rho)}$ when $\tilde{\rho}, \rho < 0$

Step 5

Putting It All Together

Infinity Norms when $\tilde{\rho}, \rho$ and $\gamma > 0$:

$$\|G^{-1}\|_\infty = \frac{1}{n(1-\rho)}\left[1+\gamma(s-2)\right]$$

$$\|\tilde{G}^{-1}\|_\infty = \frac{1}{n(1-\tilde{\rho})}\left[1+\tilde{\gamma}(s-2)\right]$$

Conclusion

Since both terms examined in the expression for G^{-1} are greater than the corresponding terms in the expression for $\tilde{G}^{-1}$ when $\tilde{\rho}, \rho$ and γ are greater than zero,

$$\|G^{-1}\|_\infty \geq \|\tilde{G}^{-1}\|_\infty$$

Furthermore, when $s = 2$ and $\tilde{\rho}, \rho$ and γ are less than zero it can be shown that the reverse is guaranteed to be true. When $s > 2$ the answer is dependent upon where $\tilde{\rho}$ or ρ fall with respect to the interval $\left[\frac{-1}{s-1}, 1\right)$. When ρ is below the lower bound but $\tilde{\rho}$ is above, the relative norm values will be ambiguous. By constraining ρ and $\tilde{\rho}$ to be greater than zero, the inequality is unambiguous.

As a result, in the positive-correlation PoC regime the binarization process *enhances positive correlation relationships* and, when $\tilde{\rho}$ and ρ are on the same side of the discontinuity at $\frac{-1}{s-1}$, it will also deemphasize negative correlation relationships (so, this will happen the vast majority of the time). Note that this seeming disadvantage can be mitigated by adding the complement to each binarized column into the feature matrix.

VITA

Jason Orender

Department of Computer Science

Old Dominion University

Norfolk, VA 23529

EDUCATION

1989-1993, B.S. Petroleum Engineering, University of Texas at Austin

2000-2002, M.B.A., George Mason University.

2015-2018, M.S. Computer Science, Old Dominion University.

2015-2026, Research Assistant, Department of Computer Science, Old Dominion University.

PROFESSIONAL EXPERIENCE

1995-2015, Nuclear Officer, U.S. Navy.

2019-2025, Data Scientist, Frontier Technology Inc.

CONFERENCE PRESENTATIONS

1. 2022 IEEE International Conference on Big Data, "LASSO Logic Engine: harnessing the logic parsing capabilities of the LASSO algorithm for longitudinal feature learning", 17-20 December 2022
2. 2025 IEEE International Conference on Big Data, "Efficient Longitudinal Feature Selection via Binarized Transformation: Theory and Case Studies", 08-11 December 2025

3. 2025 International Conference on Big Data Analytics & Applications (BDAA 2025), "Anytime Rule Compression and Rectified Logistic Modeling for Longitudinal Signals", 25-27 November 2025