Efficient Longitudinal Feature Selection via Binarized Transformation: Theory and Case Studies

Jason Orender
Department of Computer Science
Old Dominion University
Norfolk, Virginia
joren001@odu.edu

Jiangwen Sun
Department of Computer Science
Old Dominion University
Norfolk, Virginia
jsun@cs.odu.edu

Mohammed Zubair
Department of Computer Science
Old Dominion University
Norfolk, Virginia
zubair@cs.odu.edu

***Abstract*—Selecting relevant features in high-dimensional, strongly correlated longitudinal data is challenging when go/no-go thresholds and lags drive outcomes. We use binarization to map continuous measurements to binary indicators defined by critical ranges, simplifying correlations and stabilizing sparse model fitting. To avoid overstating generality, our theoretical exposition deliberately adopts the zero threshold arcsin relation as a tractable proof-of-concept: for jointly normal variables, sign binarization reduces the magnitude of pairwise correlations and yields a better-conditioned correlation matrix, which in turn improves satisfaction of the LASSO irrepresentable condition and supports recovery. More general nonzero threshold mappings obey a related but more intricate correlation formula; we therefore present the arcsin case as a conservative surrogate while implementing feature-specific critical ranges in practice. We validate the approach on synthetic data and the HAI industrial control system dataset. On HAI, the binarized model achieved accuracy 0.987, area under the receiver operating characteristic curve 0.982, F1 score 0.987, and Youden's J 0.974, while selecting only turbine-related sensors and reducing false positives. The transformation runs in linear time and produces sparse, interpretable models with actionable ranges and lags, supporting monitoring and decision making in complex systems.**

Keywords—feature selection; binarization; LASSO; longitudinal data; industrial control systems; anomaly detection; interpretability

I. Introduction

Selecting relevant predictors from high-dimensional, strongly correlated longitudinal data remains difficult, especially when outcomes are driven by thresholds and lags [1]. In such settings, standard L1-penalized methods (least absolute shrinkage and selection operator, LASSO) can return dense, unstable supports because multicollinearity saturates the correlation structure and undermines conditions required for consistent recovery [2]. This problem recurs across domains including clinical studies, industrial control systems, and finance, where delayed responses are common.

We address this by **binarizing** continuous inputs into per-feature **critical-range** indicators (values inside the range map to 1, outside to −1). This preprocessing directly **simplifies correlations**, improves conditioning of the Gram matrix, and yields sparse models with interpretable **ranges** and **lags** when combined with L1-regularized logistic regression [3]. The transformation is linear-time, making it practical at scale.

To avoid overstating generality, our theoretical exposition deliberately adopts the **zero-threshold** case as a tractable **proof-of-concept**: for jointly normal variables, sign binarization contracts pairwise correlations according to an **arcsin relation [4,5]**, which in turn bounds the inverse correlation matrix [6] and improves satisfaction of the LASSO **irrepresentable condition (IC) [2]**. In practice, we implement **feature-specific, nonzero critical ranges** whose correlation mapping is more intricate; the arcsin case is therefore presented as a conservative surrogate that captures the central mechanism driving the improvement.

Preview of results. On HAI, models fit to binarized predictors achieved the metrics above (on 2,552 training and 1,080 test examples with 590 lag-expanded features), and their coefficients correctly concentrated on the attacked turbine loop (P2), identifying the source of the failure by enhancing subsystem fidelity and interpretability [7].

A. Contributions

This paper:

- **Theory.** States lemmas connecting binarization to an arcsin-based contraction of correlations [4,5], bounding the inverse correlation matrix and tightening off-support covariances. Together, these increase the probability that IC holds and that LASSO recovers the true support [2].
- **Approach.** Describes a linear-time **critical-range binarization** pipeline (Algorithm 1) that plugs into standard L1-regularized models [3] and exposes actionable thresholds and lags.
- **Evidence.** Validates on synthetic data and on the **HAI 1.0** industrial control system (**ICS**) dataset, where binarization yields **ACC 0.987**, **AUC 0.982**, **F1 0.987**, and **Youden's J 0.974**, while selecting only turbine-related sensors and reducing false positives.

B. Related Work and Positioning

LASSO and the Irrepresentable Condition (IC). Our work builds directly on the IC of Zhao and Yu, which characterizes when LASSO can recover the true support in the presence of correlated predictors [2]. Prior art pinpoints why strong correlations derail selection but does not offer a pre-fit mechanism to reduce or bound those correlations. We inject such a mechanism through binarization: by mapping features to {±1} indicators, pairwise correlations contract (arcsin relation in the zero-threshold case) [4], the relevant Gram matrix is better conditioned [6], and the IC term is more readily satisfied. This positions binarization as a principled, model-agnostic preprocessing step rather than a new penalty or solver. (see related works below and Theory §4).

Group and structured sparsity methods. Group LASSO and related block/ordered variants address multicollinearity by imposing structure on the coefficients (e.g., group penalties or monotone lag decay) [8–12,13]. These approaches preserve domain groupings and can stabilize estimation but do not directly transform the correlation structure of the inputs. In highly correlated, lag-expanded longitudinal designs, they may remain sensitive to near-singular covariance. Our study instead tightens off-diagonal correlations before fitting, then uses standard L1 regularization to recover sparse, lag-specific supports.

Correlation-aware or rank-based LASSO variants. WLasso rewrites design matrices to account for predictor correlations, and rank-based LASSO replaces raw values by ranks to gain robustness under heavy tails [14,15]. Both can be advantageous in specific regimes but still leave the original covariance geometry largely intact or target robustness rather than IC satisfaction. By contrast, binarization systematically contracts correlations (and bounds the inverse correlation matrix), offering a conceptual route to improve the IC margin irrespective of the downstream sparse solver.

Wrapper/importance methods. Recursive feature elimination and tree-based importance (RF/GB) provide pragmatic pruning under complex dependencies but are typically heuristic, dataset-size sensitive, and less transparent about why features are retained [16]. Our pipeline yields explicit critical ranges and lags for selected sensors, which are actionable in monitoring and control contexts.

Quadratic-programming and logic-rule approaches. Katrutsa & Strijov's quadratic-programming selection and Logical Analysis of Data (LAD) aim to curb redundancy and capture thresholding via Boolean structure [17-19]. However, pairwise testing and combinatorial rule search can become computationally intensive and may struggle with groupwise dependencies among lagged, correlated signals. Our approach retains the thresholding intuition of those lines of work while achieving polynomial complexity by pairing binarization with L1-regularized logistic regression.

Scope decisions shaped by related work. In setting scope and comparisons, we (i) treat Weighted Lasso [20] as complementary rather than a head-to-head competitor; (ii) exclude deep learning or manifold-projection feature-selection methods where transparency and efficiency are not the primary design goals; and (iii) acknowledge rule/rough-set families but do not emphasize them experimentally due to their scaling characteristics on large, lag-expanded designs. These choices reflect our focus on a theoretically motivated, computationally light preprocessing step that improves IC satisfaction and interpretability for longitudinal systems. (see Introduction §1).

Impact on this paper's claims and design. Because nonzero thresholds yield more intricate correlation formulas, we present the zero-threshold arcsin relation as a *proof-of-concept* that conservatively captures the contraction mechanism; in practice, we implement feature-specific critical ranges and show empirically that the benefits persist. This positioning avoids overstating generality while still linking the method to well-understood correlation bounds that inform our theory, experiment design, and interpretation of results. (see Theory §4).

C. Organization

Section II reviews background and related work; Section III details the binarization pipeline approach; Section IV presents the theoretical results; Section V reports experiments with the case studies; Section VI discusses implications and limitations; and Section VII concludes.

II. Background

This section summarizes the statistical background that motivates a binarization-first pipeline for sparse modeling in longitudinal, highly correlated settings. We review the L1-regularized modeling setup [3], formalize the irrepresentable condition (IC) for support recovery [2], and state the correlation-contraction property induced by zero-threshold sign binarization. We use the arcsin relation as a tractable proof-of-concept (PoC) case to avoid overstating generality [4,5]; nonzero thresholds obey a related but more intricate mapping that we do not rely on for the theory presented here.

A. Problem Setting and Notation

Let

$$X \in \mathbb{R}^{n\times d}$$

denote standardized predictors (zero mean, unit variance) constructed from longitudinal streams and their lags;

$$y \in \{0,1\}$$

is the response. The true support is

$$S \subset \{1, \dots, d\}$$

with:

$$|S| = s, \text{ and } S^c \text{ its complement}$$

For any index set A, write X_A for the corresponding submatrix. The (sample) Gram matrix on S is $G = \frac{1}{n} X_S^\top X_S$. For the binarized design $\tilde{X} \in \{\pm 1\}^{n\times d}$, define $\tilde{G} = \frac{1}{n} \tilde{X}_S^\top \tilde{X}_S$. We use ρ for pairwise correlations in the continuous domain and r for those after binarization (see Theory §4.1 for matching notation).

B. L1-regularized Modeling and Support Recovery

Throughout, we fit a sparse linear or logistic model with an ℓ_1 penalty:

$$(\hat{\beta}_0, \hat{\beta}) = \underset{\beta_0,\beta}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} \ell\,(y_i,\ \beta_0 + x_i^{\top}\beta) + \lambda \parallel \beta \parallel_1 \quad (1)$$

where ℓ is the squared-error or logistic loss [3]. The LASSO's ability to recover S is characterized by the irrepresentable condition (IC), which requires the off-support predictors not be too correlated with the on-support set [2]:

$$\|X_{S^c}^{\top} X_S\,(X_S^{\top} X_S)^{-1}\mathrm{sign}(\beta_S)\|_{\infty} < 1 \quad (2)$$

High collinearity among raw, lag-expanded features often violates (2), yielding dense or unstable selections. The core idea behind our approach is to alter the correlation geometry before fitting so that the left-hand side of (2) shrinks, improving the chance of correct support recovery. (See Theory §4 for the IC-based development.)

C. *Zero-threshold Binarization and the Arcsin Relation (PoC)*

As a proof-of-concept theoretical setting, consider sign binarization at zero threshold under joint normality of the continuous variables. Let (X, Y) be standardized, jointly normal with Pearson correlation ρ, and define $\tilde{X} = \mathrm{sign}(X) \in \{\pm 1\}$, $\tilde{Y} = \mathrm{sign}(Y)$. Then the correlation between the binarized variables satisfies the classical arcsin relation:

$$\tilde{\rho} \ = \ \frac{2}{\pi}\arcsin(\rho), \qquad |\tilde{\rho}| \le |\rho| \text{ for } \rho \quad (3)$$

$\in (-1,1)$, with strict inequality except at $0, \pm 1$.

This orthant-probability result (Hotelling-type) shows that sign binarization contracts correlation magnitudes and, in turn, avoids saturation of off-diagonals in $\tilde{G}$ [4,5]. We leverage (3) in our lemmas and theorem as a conservative surrogate for the more general, but algebraically more involved, nonzero-threshold mappings used in practice.

D. *Matrix-conditioning consequences*

Correlation contraction has matrix-level implications. In the constant-correlation stylized case with on-support pairwise correlation ρ (continuous) and r (binarized), we have:

$$G = n\big((1-\rho)I_s + \rho J_s\big),$$
$$G^{-1} = \frac{1}{n(1-\rho)}\left(I_s - \frac{\rho}{1+\rho(s-1)}J_s\right) \quad (4)$$

and analogously $\tilde{G}$ with r in place of ρ. When $0 < \tilde{\rho} < \rho < 1$, row-sum bounds imply:

$$\parallel \tilde{G}^{-1} \parallel_{\infty} \le \parallel G^{-1} \parallel_{\infty}, \quad (5)$$

so the binarized Gram matrix is no worse conditioned, and often strictly better, than its continuous counterpart [6] (see also the explicit ∞-norm contraction via row-sum/diagonal-dominance at the end of Lemma 2). Together with the off-support covariance contraction:

$$\parallel \tilde{C}_i \parallel_{\infty} \le \parallel C_i \parallel_{\infty},$$
$$C_i = 1/n\, X_{i,S^c}^{\top} X_S, \quad (6)$$
$$\tilde{C}_i = 1/n\, \tilde{X}_{i,S^c}^{\top} \tilde{X}_S,$$

these bounds shrink the IC term for each $i \in S^c$:

$$\tilde{\theta}_i = \tilde{C}_i\, \tilde{G}^{-1}\mathrm{sign}(\beta_S) \quad \text{vs.} \quad \theta_i = C_i\, G^{-1}\mathrm{sign}(\beta_S), \quad (7)$$

increasing the probability that $\parallel \tilde{\theta}_i \parallel_{\infty} < 1$. (Lemmas 2-4 show the Sherman-Morrison derivation and norm comparison [6].)

E. *Practical bridge to the approach*

While our theory section uses the zero-threshold arcsin case as a clean PoC, our approach implements feature-specific critical ranges: a value maps to +1 if it lies within the empirically derived range observed during events and to -1 otherwise (Algorithm 1). This maintains the intuition of correlation contraction and the benefits for IC satisfaction, while yielding directly interpretable thresholds and lags in longitudinal applications. (Approach §3 and Algorithm 1).

III. APPROACH

This section presents a practical, scalable pipeline that (i) binarizes longitudinal predictors into feature-specific critical-range indicators, and (ii) fits a sparse L1-regularized classifier on the transformed design. The binarization step is a single pass over the data, producing interpretable thresholds and revealing lags while simplifying the correlation geometry for LASSO. We emphasize that while our theory uses the zero-threshold arcsin case as a proof-of-concept, the implementation here estimates nonzero, per-feature critical ranges from the positive class and leverages them in practice [4,5].

A. *Critical-Range Binarization*

Let $X \in \mathbb{R}^{n\times d}$ be the lag-expanded design (Section II-A) and $y \in \{0,1\}$ the response. Define the event subset

$$Z \ = \ \{\, i \in \{1, \dots, n\} \colon y_i = 1\,\} \quad (8)$$

and for each feature $j \in \{1, \dots, d\}$ compute its critical range (CR) from event rows only:

$$\big[b_j,\, c_j\big] \ = \ \Big[\min_{i\in Z} x_{ij},\ \max_{i\in Z} x_{ij}\Big] \quad (9)$$

We then produce the binarized matrix $\tilde{X} \in \{\pm 1\}^{n\times d}$ feature-wise via:

$$\tilde{x}_{ij} \ = \ \begin{cases} +1, & x_{ij} \in \big[b_j,\, c_j\big], \\ -1, & \text{otherwise.} \end{cases} \quad (10)$$

Equations (8)-(10) implement the critical-range indicator used throughout our experiments and figures; this is the same construction summarized as Algorithm 1.

Lag expansion. For p raw streams and L historical steps, we form $d = p(L+1)$ columns by concatenating per-stream lags $\{0, \dots, L\}$ before applying (8)–(10). This preserves **lag interpretability** in $\tilde{X}$: selected columns map directly to stream-lag pairs (sensor,lag) and their CRs $\big[b_j, c_j\big]$.

Sign conventions. We use $\{\pm 1\}$ coding for alignment with the theory; a $\{0,1\}$ variant can be substituted with no change in the downstream fitting step. For domains dominated by negative-correlation mechanisms, adding **complement columns** for selected features can help mirror "outside-range" activation (see Appendix remark).

Algorithm 1 Binarization

```
Require: $\mathbf{X}$, $y$ (training set only)
Output: $\mathbf{A}$, $r_j[b, c]$
Let $r_j[b, c]$ be the critical range for feature vector $\mathbf{X}_j$, where $b$ is the minimum and $c$ is maximum defining that range.
procedure BINARIZATION($\mathbf{X}$, y)
    Let Z be the rows of $\mathbf{X}$ where $y$=1.
     $Z \subseteq \mathbf{X}$, where $Z = \{\mathbf{X} \mid y = \mathbf{TRUE}\}$
    Compute $r_{j=1..d} = [min(Z_j), max(Z_j)]$ .
    Note: If the min / max calculations induce brittleness in the solution (e.g. if there are outliers in the data set), using a robust quantile can provide a remedy.
    Compute $\mathbf{A}$, the set of transformed features, such that:
        $a_{ij} = +1$ when $x_{ij} \geq r_j[b]$ and $x_{ij} \leq r_j[c]$
        $a_{ij} = -1$ when $x_{ij} < r_j[b]$ or $x_{ij} > r_j[c]$
end procedure
```

***Important**: The critical ranges are only calculated using the training set, not the set-aside test set. The critical ranges are retained and used again when processing the test set. This prevents information leakage and maintains the integrity of the data.*

B. Model fitting on Binarized Features

We fit a sparse linear or logistic model with an ℓ_1 penalty on $\tilde{X}$ for $(\hat{\beta}_0, \hat{\beta})$ similar to eq. 1:

$$\underset{\beta_0, \beta}{\operatorname{argmin}} \ \frac{1}{n} \sum_{i=1}^{n} \ell\,(y_i,\ \beta_0 + \tilde{x}_i^{\top} \beta) + \lambda \parallel \beta \parallel_1 \qquad (11)$$

selecting λ by K-fold cross-validation (we use $K = 10$) via standard coordinate-descent solvers (e.g., glmnet, choosing $\lambda_{\min}$ or the 1-SE rule). This pairing is model-agnostic and leverages mature, efficient implementations.

C. Complexity and Scalability

Computing $[b_j, c_j]$ and applying the thresholding each require a single pass per feature. Thus the transform is linear-time in the number of feature values:

$$\text{Time} = \Theta(nd), \qquad \text{Space} = \Theta(nd) \qquad (12)$$

and is trivially parallelizable across features. In our experiments this overhead is negligible relative to the solver, and empirically, even including transformation time, fits on $\tilde{X}$ are often faster than fits on raw X.

D. Outputs and Interpretability Artifacts

The pipeline yields three artifacts per selected column j:

1. Binary indicator $\tilde{x}_{\cdot j}$ exposing when the feature is "active."
2. Critical range $[b_j, c_j]$, which is directly actionable (thresholds for monitoring/alerting).
3. Lag index (from the column's construction), enabling attribution of when the feature matters.

These artifacts underpin the qualitative analyses in Section V (e.g., turbine-only selection on HAI and reduced false positives), and they are central to the claim of **interpretable sparsity**.

E. Practical Notes

- **Zero-threshold vs. nonzero thresholds**. Our theory uses the zero-threshold arcsin mapping as a conservative, tractable proxy; the implementation uses nonzero, feature-specific CRs learned from events. This separation avoids overstating generality while retaining the core contraction mechanism that benefits LASSO's IC [4,5].
- **Baselines and CV protocol**. We compare $\tilde{X}$ to raw X under the same solver and CV protocol; additional baselines (e.g., RF) can be included for context.

IV. Theory

This section develops the theoretical core showing how binarization improves the likelihood that an ℓ_1-regularized model satisfies the irrepresentable condition (IC) and recovers the true support [2]. We deliberately use the zero-threshold, jointly normal setting as a tractable proof-of-concept (PoC): in this regime, the arcsin relation yields an exact contraction of pairwise correlations after sign/binary mapping [4,5], which tightens off-support covariances and improves Gram-matrix conditioning [6]. The more general nonzero-threshold mappings used in our implementation obey related but more intricate formulas; we therefore present the arcsin case as a conservative surrogate that captures the key mechanism.

A. Setup and Assumptions (PoC case)

Let $X \in \mathbb{R}^{n \times d}$ be the standardized continuous design; $y \in \{0,1\}$ is the response; $S \subset \{1, \dots, d\}$ is the true support, $|S| = s$, with complement S^c. Write X_S and X_{S^c} for submatrices and define the (sample) Gram matrix $G = \frac{1}{n} X_S^{\top} X_S$. Let $\tilde{X} \in \{\pm 1\}^{n \times d}$ be the binarized design (zero-threshold sign mapping for theory), with $\tilde{G} = \frac{1}{n} \tilde{X}_S^{\top} \tilde{X}_S$. Denote by ρ the pre-binarization Pearson correlation and by r the post-binarization correlation.

Assumptions (PoC). (i) Columns of X are standardized; (ii) joint normality of the continuous features; (iii) theory uses **zero threshold** sign binarization, while practice uses **feature-specific** (nonzero) critical ranges; (iv) large-n approximations replace sample covariances with population quantities where stated.

B. Lemma 1 (Arcsin contraction of correlation)

Statement. For standardized, jointly normal (X, Y) with correlation ρ, and binarized $\tilde{X} = \operatorname{sign}(X)$, $\tilde{Y} = \operatorname{sign}(Y)$ (or equivalently 0/1 with centering), the correlation after binarization satisfies:

$$\tilde{\rho} \;=\; \frac{2}{\pi}\arcsin(\rho), \text{ hence } |\tilde{\rho}| \le |\rho| \text{ for } \rho \in (-1,1) \tag{13}$$

with strict inequality except at $\rho \in \{0, \pm 1\}$ [4,5].

Proof sketch. Using orthant probability of the bivariate normal, $P(X > 0, Y > 0) = 1/4 + 1/2\pi \arcsin(\rho)$; with $E[\tilde{X}] = E[\tilde{Y}] = 0$ after centering, $\operatorname{Corr}(\tilde{X}, \tilde{Y}) = 4(P(X > 0, Y > 0) - 1/4) = 2/\pi \arcsin(\rho)$. Convexity of arcsin on $[0,1]$ and symmetry yield $|\tilde{\rho}| \le |\rho|$ [4,5].

C. Lemma 2 (Bound on the inverse correlation/Gram matrix)

Statement. If the post-binarization correlation matrix $R = E[\tilde{X}^{\top}\tilde{X}]$ is invertible, then its smallest eigenvalue is bounded away from 0, implying $\| R^{-1} \|_2 \le \gamma$ for some finite γ. Consequently, for the on-support Gram matrix $\tilde{G}$, $\| \tilde{G}^{-1} \|$ is no worse, and often strictly better, than $\| G^{-1} \|$.

Sketch & intuition. Lemma 1 keeps all off-diagonal entries of R strictly away from ± 1 (unless a column is a scalar multiple), ensuring positive definiteness and a finite spectral-norm bound $\| R^{-1} \|_2 \le 1/\lambda_{\min}(R)$. A stylized constant-correlation model makes this explicit:

$$\begin{gathered} G = n\big((1-\rho)I_s + \rho J_s\big), \\ G^{-1} = \frac{1}{n(1-\rho)}\left(I_s - \frac{\rho}{1+\rho(s-1)}J_s\right) \end{gathered} \tag{14}$$

and analogously for $\tilde{G}$ with ρ replaced by $\tilde{\rho}$. Since $0 < \tilde{\rho} < \rho < 1$ under Lemma 1, row-sum bounds yield $\| \tilde{G}^{-1} \|_\infty \le \| G^{-1} \|_\infty$ [6].

Finally, by a standard row-sum/diagonal-dominance bound (e.g., Gershgorin/Varah), writing $s_i = \sum_{k \ne i} |R_{ik}|$ and $s_\star = \max_i s_i$, we have $\| R^{-1} \|_\infty \le (1 - s_\star)^{-1}$; since zero threshold binarization contracts off-diagonals entrywise, $\tilde{s}_\star \le s_\star$ and thus $\| \tilde{R}^{-1} \|_\infty \le (1 - \tilde{s}_\star)^{-1} \le (1 - s_\star)^{-1}$. Since G_{SS} is the on-support block of the Gram/correlation matrix, the same ∞-norm **upper bound** applies to G_{SS}^{-1} and is no larger after binarization.

D. Lemma 3 (Contraction of off-support covariances)

Statement. For $i \in S^c$ and $j \in S$, the magnitude of covariance decreases under binarization:

$$\big(\tilde{x}_i, \tilde{x}_j\big)| \;\le\; |\operatorname{Cov}\big(x_i, x_j\big)| \tag{15}$$

Sketch. With the arcsin relation applied to $\big(x_i, x_j\big)$ at zero threshold, $\operatorname{Cov}\big(\tilde{x}_i, \tilde{x}_j\big) = 1/2\pi \arcsin\big(\rho_{ij}\big)$ (from the orthant-probability already cited), hence $\big|\operatorname{Cov}\big(\tilde{x}_i, \tilde{x}_j\big)\big| \le \big|\rho_{ij}\big| = \big|\operatorname{Cov}\big(x_i, x_j\big)\big|$ [4,5].

E. Lemma 4 (IC bound tightening under binarization)

Let S be the true support with $|S| = s$ and S^c its complement (definition repeated for clarity).
For any $j \in S^c$, write the IC component as:

$$\begin{aligned} \theta_j \;&=\; C_{jS}\, G_{SS}^{-1} \operatorname{sign}\big(\beta_S\big), \\ \tilde{\theta}_j \;&=\; \tilde{C}_{jS}\, \tilde{G}_{SS}^{-1} \operatorname{sign}\big(\beta_S\big) \end{aligned} \tag{16}$$

where C_{jS} and $\tilde{C}_{jS}$ denote the off–support/on–support covariance row (pre- and post-binarization), and G_{SS}, $\tilde{G}_{SS}$ are the on-support Gram blocks.

Statement. Under Lemmas 1–3 (pairwise correlation contraction; inverse-Gram bound; off-support covariance contraction) and submultiplicativity of induced norms,

$$\begin{aligned} \| \theta_j \|_\infty &\le \| C_{jS} \|_\infty \| G_{SS}^{-1} \|_\infty \| \operatorname{sign}(\beta_S) \|_\infty, \\ \| \tilde{\theta}_j \|_\infty &\le \| \tilde{C}_{jS} \|_\infty \| \tilde{G}_{SS}^{-1} \|_\infty \| \operatorname{sign}(\beta_S) \|_\infty. \end{aligned} \tag{17}$$

Since $\| \operatorname{sign}(\beta_S) \|_\infty = 1$, define the per-index bounds:

$$\begin{aligned} B_j \;&:= \| C_{jS} \|_\infty \| G_{SS}^{-1} \|_\infty, \\ \tilde{B}_j \;&: = \| \tilde{C}_{jS} \|_\infty \| \tilde{G}_{SS}^{-1} \|_\infty \end{aligned} \tag{18}$$

Then Lemmas 1–3 yield the bound contraction:

$$\tilde{B}_j \;\le\; B_j \tag{19}$$

and hence,

$$\begin{gathered} \| \tilde{\theta}_j \|_\infty \;\le\; \tilde{B}_j \;\le\; B_j, \\ \| \theta_j \|_\infty \;\le\; B_j \end{gathered} \tag{20}$$

for all $j \in S^c$.

So, if $B^\star : = \max_{j \in S^c} B_j < 1$, then $\| \theta \|_\infty < 1$ (the IC holds). After binarization,

$$\tilde{B}^\star : = \max_{j \in S^c} \tilde{B}_j \;\le\; \max_{j \in S^c} B_j \;=\; B^\star \tag{21}$$

so the norm-based sufficient condition $\tilde{B}^\star < 1$ is no harder, and often easier, to satisfy.

This lemma establishes tightening of a standard upper bound on each IC component (and on its maximum). It does not assert a pointwise ordering $\| \tilde{\theta}_j \|_\infty \le \| \theta_j \|_\infty$ in every dataset; rather, it shows that the admissible upper envelope shrinks under binarization, which increases the chance that the IC is met.

F. Corollary (equicorrelation with aligned signs)

Let S be the true support with $|S| = s$, and fix any $j \in S^c$. Assume the stylized constant-correlation model used in §II.D: for standardized columns, $\operatorname{Corr}(X_k, X_\ell) = \rho$ for $k \ne \ell \in S$ and $\operatorname{Corr}\big(X_j, X_k\big) = \rho$ for $k \in S$, with $0 < \rho < 1$.

Without loss of generality, absorb signs into the columns so that $\operatorname{sign}(\beta_S) = 1$. Write the population correlation blocks as:

$$\begin{gathered} \Sigma_{SS} = (1-\rho)I_s + \rho\, \mathbf{1}\mathbf{1}^{\top}, \\ \Sigma_{jS} = \rho\, \mathbf{1}^{\top} \end{gathered} \tag{22}$$

Then the IC scalar (cf. §IV.E) reduces to:

$$\theta_j \;=\; \Sigma_{jS}\, \Sigma_{SS}^{-1} \operatorname{sign}(\beta_S) \;=\; \rho\, \mathbf{1}^{\top} \Sigma_{SS}^{-1} \mathbf{1} \tag{23}$$

Using the equicorrelation inverse,

$$\Sigma_{SS}^{-1} = \frac{1}{1-\rho} I_s \; - \; \frac{\rho}{(1-\rho)(1-\rho+s\rho)} \mathbf{11}^{\top} \tag{24}$$

we obtain:

$$\begin{aligned}\mathbf{1}^{\top}\Sigma_{SS}^{-1}\mathbf{1} = \frac{s}{1-\rho+s\rho} \quad &\Rightarrow \quad \theta_j \; = \; \frac{s\rho}{1-\rho+s\rho} \\ &= \; \frac{s\rho}{1+(s-1)\rho}\end{aligned} \tag{25}$$

After binarization at zero threshold (PoC setting), correlations contract under the same conditions as (5):

$$\begin{gathered}\tilde{\rho} \; = \; \frac{2}{\pi}\arcsin(\rho), \\ 0 < \tilde{\rho} < \rho < 1,\end{gathered} \tag{26}$$

and the same calculation gives:

$$\tilde{\theta}_j \; = \; \frac{s\tilde{\rho}}{1+(s-1)\tilde{\rho}} \tag{27}$$

Define $f(x) = \frac{sx}{1+(s-1)x}$ on $(-1/(s-1), 1)$. Then,

$$f'(x) \; = \; \frac{s}{(1+(s-1)x)^2} \; > \; 0 \tag{28}$$

so f is strictly increasing. Because $0 < \tilde{\rho} < \rho < 1$, we have:

$$\left|\tilde{\theta}_j\right| \; = \; f(\tilde{\rho}) \; < \; f(\rho) \; = \; \left|\theta_j\right| \tag{29}$$

Hence, in this stylized equicorrelation regime with aligned signs,

$$\| \, \tilde{\theta}_j \, \|_\infty \; = \; \left|\tilde{\theta}_j\right| \; < \; \left|\theta_j\right| \; = \| \, \theta_j \, \|_\infty, \tag{30}$$

i.e., the IC term for each $j \in S^c$ is **strictly smaller after binarization**, supporting the more general conclusion.

G. Main Theorem (Higher probability that a standard IC sufficient condition holds).

Let B_j and $\tilde{B}_j$ be as in Lemma 4:

$$\begin{aligned} B_j &:= \| \, C_{jS} \, \|_\infty \, \| \, G_{SS}^{-1} \, \|_\infty, \\ \tilde{B}_j &: = \| \, \tilde{C}_{jS} \, \|_\infty \, \| \, \tilde{G}_{SS}^{-1} \, \|_\infty \end{aligned} \tag{31}$$

and define $B^\star : = \max_{j \in S^c} B_j$ and $\tilde{B}^\star : = \max_{j \in S^c} \tilde{B}_j$.

Statement. Under the PoC assumptions (zero threshold sign binarization under joint normality) and large-n approximations that replace sample covariances by their population limits, Lemmas 1–4 imply $\tilde{B}^\star \leq B^\star$. Consequently,

$$\Pr\left(\tilde{B}^\star < 1\right) \; \geq \; \Pr(B^\star < 1) \tag{32}$$

Since $\{\tilde{B}^\star < 1\} \subseteq \{\| \, \tilde{\theta} \, \|_\infty < 1\}$ and $\{B^\star < 1\} \subseteq \{\| \, \theta \, \|_\infty < 1\}$, binarization weakly increases the probability of satisfying a standard sufficient condition for the IC.

Sketch. Lemma 1 (arcsin contraction) yields $\| \, \tilde{C}_{jS} \, \|_\infty \leq \| \, C_{jS} \, \|_\infty$; Lemma 2 (inverse-Gram bound) gives $\| \, \tilde{G}_{SS}^{-1} \, \|_\infty \leq \| \, G_{SS}^{-1} \, \|_\infty$ (e.g., via row-sum/diagonal-dominance bounds or in the constant-correlation stylized case). Hence $\tilde{B}_j \leq B_j$ for all $j \in S^c$ and thus $\tilde{B}^\star \leq B^\star$. The event inclusions follow from $\| \, \theta_j \, \|_\infty \leq B_j$ and $\| \, \tilde{\theta}_j \, \|_\infty \leq \tilde{B}_j$. No general pointwise ordering between $\| \, \tilde{\theta} \, \|_\infty$ and $\| \, \theta \, \|_\infty$ is claimed; the equicorrelation corollary provides a strict ordering only in that stylized regime.

H. Practical significance, caveats, and implementation bridge

Significance. The lemmas isolate why binarization helps: (i) pairwise contraction keeps off-diagonals from saturating; (ii) better conditioning bounds inverse norms; (iii) smaller off-support covariances reduce the IC term. Together these raise the chance of correct support recovery and promote sparse, stable selections.

Negative correlations & discontinuity. In the constant-correlation model, row-sum bounds for $\| \, \tilde{G}^{-1} \, \|_\infty$ are unambiguous for $0 < r < \rho < 1$, while for strong negative correlations a discontinuity at $-1/(s-1)$ complicates monotonicity; in practice, adding complement columns (the negative of those columns) for selected features mitigates adverse cases.

Bridge to the approach. Our implementation uses feature-specific, nonzero critical ranges (Algorithm 1) to produce $\tilde{X} \in \{\pm 1\}^{n \times d}$; while their exact correlation mapping is more involved, empirical results and the PoC chain above explain why the same mechanism (correlation tightening ⇒ improved IC margin) appears in practice (see Approach §3 for Algorithm 1).

V. Results

This section reports empirical evidence that the binarization-first pipeline improves accuracy, sparsity, and interpretability across synthetic and real-world settings. We adhere to a common protocol: lag expansion, critical-range binarization (Algorithm 1), and L1-regularized logistic regression with 10-fold cross-validation; comparative baselines mirror the same CV setup.

A. Synthetic Data Experiments

Design. We generated longitudinal data with 7 relevant and 33 irrelevant variables over 4911 examples. Exactly one lag per relevant variable triggers the event, yielding 11 time steps and 440 features (≈ 1.6 % relevant). This setting stresses support recovery under strong correlation and lag structure (details and figures in §5).

Findings. LASSO on **transformed (binarized)** features nearly perfectly classifies the test set and returns a **sparse, lag-faithful** support; untransformed fits are noisier, denser, and mis-localized (Fig. 1a-b). A quadratic-programming selector (Katrutsa–Strijov) recovers at most one true feature (see Fig. 2) and is **100–200**× slower than binarization plus LASSO.

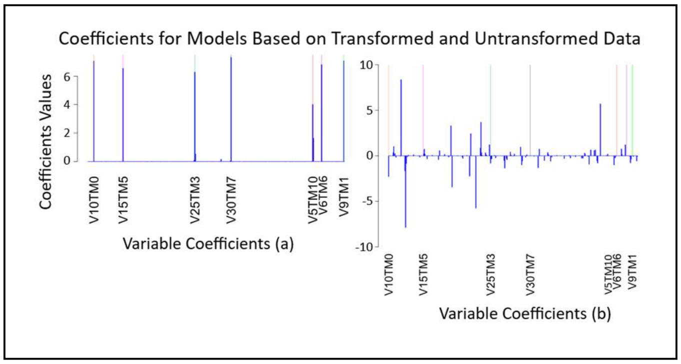

Fig. 1. Coefficients for transformed (a) vs. untransformed (b) data. Adapted from [21].

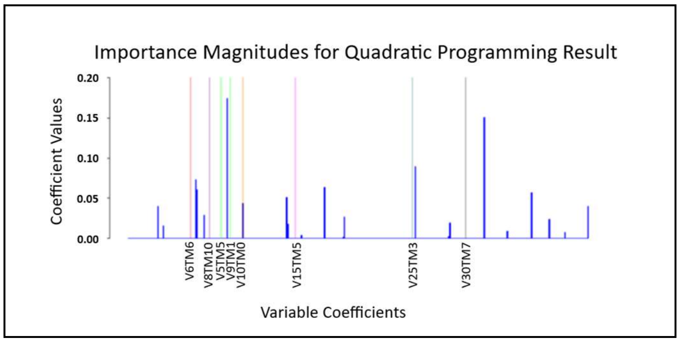

Fig. 2. Katrutsa-Strijov coefficient response on untransformed data. Adapted from [21].

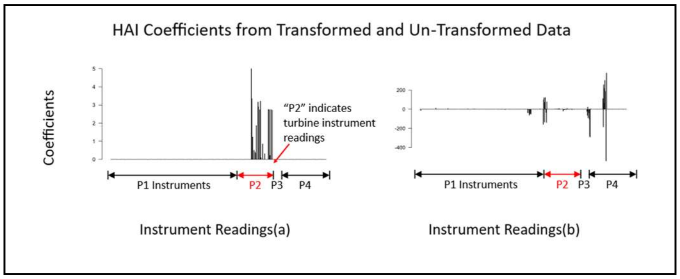

Fig. 3. HAI Coefficients for transformed (a) and untransformed (b) data. Adapted from [21]. Feature groups are shown along the x-axis and separated into four parts (1-4). Each bar represents the coefficient for a particular feature at a particular lag (590 total).

Takeaway. Binarization improves accuracy, halves compute (despite transform cost – See Table 1), and most importantly, exposes the correct lagged support under severe multicollinearity.

TABLE I. SYNTHETIC CASE COMPARISON ON UNSEEN TEST DATA

Method	Metrics			
	Transform	***Rel. Time***	***Youden's J***	***F1***
LASSO	Transformed	1.00	0.975	0.988
LASSO	Untransformed	1.83	0.753	0.804
Ridge*	Transformed	1.00	0.980	0.990
Ridge	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

** Small loss of lag fidelity noted for Ridge when indicating specific time steps (journal note under Table 1).*

B. HAI ICS Case Study

Dataset & setup. The HAI 1.0 ICS testbed spans boiler, turbine, and water treatment loops with real attack scenarios. We use the turbine-attack subset with 2,552 training examples, 1,080 test examples, and 59 raw sensors × 10 lags = 590 features.

Metrics. We report Accuracy (ACC), AUC, F1, and Youden's *J* on the held-out test set (operating point from CV).

TABLE II. HAI TEST-SET PERFORMANCE (TURBINE ATTACK)

Statistic	HAI Metrics	
	Transformed	***Untransformed***
ACC	0.987	0.939
AUC	0.982	0.943
F1	0.987	0.942
Youden's J	0.974	0.878

Interpretation. Results show markedly fewer false positives for the transformed model. Coefficient stems concentrate on turbine (P2) sensors only, evidencing subsystem-faithful selection; the untransformed model spreads small weights across non-attacked loops (Fig. 3a-b).

C. Additional Experiments

We include summaries of two further datasets to indicate broader evaluation without extensive detail here.

- **Goose Bay Ionosphere (UCI).** 34 features (17 pulses × real/imag). Transformed LASSO outperforms untransformed and approaches original neural-network performance; the sparse coefficient vector pinpoints pulses and yields critical energy ranges.
- **UNICEF longitudinal (1256 examples × 316 features, over 75 years).** Conventional LASSO on un-transformed features posts higher raw metrics, but at the cost of a dense, harder-to-interpret coefficient profile; transformed models are much sparser and more transparent.

TABLE III. OTHER EXPERIMENTS AT A GLANCE

Dataset	Metrics		
	ACC (Transformed / Untransformed)	***F1 (Transformed/ Untransformed)***	***Youden's J (Transformed / Untransformed)***
Goose Bay Ionosphere	0.940 / 0.871	0.993 / 0.903	0.946 / 0.866
UNICEF Longitudinal	0.902 / 0.963	0.894 / 0.958	0.815 / 0.925

D. Aggregate Observations

Across synthetic and ICS data, **critical-range binarization** consistently sharpens support recovery and reduces false positives, in line with the theoretical contraction of correlations and improved IC margin. Even where untransformed metrics win (UNICEF), binarization delivers interpretable sparsity and actionable ranges, which is often the decisive requirement in monitoring and controls.

VI. Discussion and Limitations

This section interprets the theoretical and empirical findings, clarifies the scope of claims, and delineates limitations and edge cases. We emphasize that our theoretical analysis adopts the **zero threshold** arcsin relation as a tractable proof-of-concept, while the **implemented** binarization uses **feature-specific, nonzero critical ranges**. The latter follows a more intricate correlation mapping, but the same contraction mechanism explains the empirical gains we observe.

A. When and why binarization helps

Thresholded, lagged phenomena with correlated predictors. In domains where events are triggered when variables enter critical ranges, often with delays, continuous measurements create saturated pairwise correlations that impede sparse recovery. Mapping to $\{\pm 1\}$ indicators reduces correlation magnitudes (arcsin contraction in the zero threshold, jointly normal case), improves conditioning, and lowers off-support covariances. In our experiments, this manifests as sharper supports and fewer false positives (e.g., HAI turbine loop), consistent with the theoretical chain culminating in a smaller IC term.

Interpretability and actionability. Each selected column yields a threshold range $[b_j, c_j]$ and a lag index, producing compact, actionable rules ("when sensor j is in range at lag ℓ"). This interpretability underpins the subsystem-faithful selection seen in HAI and the pulse-specific findings in the ionosphere data.

Computational efficiency. The transformation is linear-time in the number of feature values (single pass per feature), parallelizable across columns and empirically yields fits that are faster than modeling the raw design, even after including transform time (≈1.8× speedup vs. untransformed in the synthetic case – See Table 1).

B. What the theory guarantees, and what it does not

Proof-of-concept regime. The arcsin mapping $\tilde{\rho} = 2/\pi \arcsin(\rho)$ and the corollaries $|\tilde{\rho}| \leq |\rho|$ are invoked under joint normality and zero thresholds. These yield: (i) pairwise contraction; (ii) inverse-Gram bounds (better or no-worse conditioning); and (iii) reduced off-support covariances, together implying that the IC term shrinks:

$$\begin{gathered} \| \tilde{\theta}_j \|_\infty \leq \tilde{B}_j \leq B_j, \\ \| \theta_j \|_\infty \leq B_j \end{gathered} \tag{33}$$

Thus, the probability that the IC holds is weakly higher after binarization in this PoC setting. We do not claim a closed-form mapping for general nonzero thresholds; instead, we use the arcsin case as a conservative surrogate for the contraction mechanism observed empirically.

Matrix-level caveats. In the constant-correlation model, improvements are unambiguous for $0 < \tilde{\rho} < \rho < 1$. With strong negative correlations, a discontinuity at $-1/(s-1)$ can complicate monotonicity of $\| \tilde{G}^{-1} \|_\infty$ vs. $\| G^{-1} \|_\infty$. In practice, adding complement columns (to capture "outside-range" activation explicitly) mitigates adverse cases.

C. Empirical scope: what improved and where

Synthetic and HAI datasets. Binarization delivered large gains: near-perfect classification and correct lag localization on synthetic data, and ACC 0.987/AUC 0.982/F1 0.987/J 0.974 on HAI with turbine-only supports and fewer false positives than untransformed models (See Table 2 and Fig. 3). These patterns align with the theory's correlation-contraction narrative and the IC margin improvement.

UNICEF longitudinal data. A counterexample to blanket dominance: untransformed models produced higher raw metrics (e.g., F1 0.958 vs. 0.894 – See Table 3) but at the cost of dense, hard-to-interpret coefficients. The transformed model remained much sparser while still describing nearly all of the data variation, preserving attribution clarity despite lower headline scores, consistent with the method's linear decision boundary and possible information loss from binarization in complex, non-threshold regimes.

D. Limitations

1. **Linearity of the downstream model.** Our pipeline pairs binarization with **L1-regularized linear/logistic models**. Where decision boundaries are **strongly nonlinear** and not well approximated by logical combinations of range indicators, performance can lag, as seen with UNICEF.

2. **Assumptions behind the PoC theory.** The arcsin derivations assume **joint normality** and **zero thresholds**; real systems use **nonzero range** thresholds. We explicitly avoid overstating generality by presenting the arcsin case as a **proof-of-concept** rather than a universal mapping.

3. **Negative-correlation edge cases.** Around the $-1/(s-1)$ discontinuity (constant-correlation stylization), inverse-norm comparisons can become ambiguous, and monotonic improvements are **not guaranteed**. Complement features alleviate—but do not theoretically eliminate—this concern.

4. **Comparative breadth.** We benchmarked against several methods in the synthetic study and focused on **before/after** comparisons elsewhere for brevity. Some families (e.g., **rough/fuzzy-rough sets**, certain **deep** selectors) were **de-emphasized** due to computational cost or misalignment with the paper's goals of **interpretability** and **efficiency**; WLasso is

positioned as complementary rather than a head-to-head competitor.

E. Practical guidance and mitigations

- Design for interpretability. Prefer per-feature critical ranges that produce compact supports and subsystem faithfulness; report $[b_j, c_j]$ and lag indices with the selected features to enable operational use (alerts, controls).
- Handle negative correlations. Include complement columns for features likely to activate outside the event range; this can stabilize selection when relevant signals are negatively associated.
- Scale easily. Leverage the $O(nd)$ transform and parallelize across features; in many workloads, end-to-end time is lower than modeling raw X.

VII. Conclusion

We presented a binarization-first pipeline for longitudinal feature selection that (i) transforms continuous measurements into feature-specific critical-range indicators and (ii) fits a sparse ℓ_1-regularized model on the transformed design. The theoretical core, deliberately framed as a proof-of-concept, uses the zero threshold, jointly normal setting to make the argument transparent: sign binarization contracts pairwise correlations via the arcsin relation, $r = 2/\pi \arcsin(\rho)$, which bounds inverse-Gram norms and reduces off-support covariances. This chain tightens the irrepresentable condition (IC) term, yielding $\| \tilde{\theta}_j \|_\infty \leq \tilde{B}_j \leq B_j$, and $\| \theta_j \|_\infty \leq B_j$, thereby increasing the likelihood of correct support recovery. We emphasize that this arcsin analysis is a tractable surrogate; in practice we implement nonzero, per-feature ranges, whose correlation mapping is more intricate but empirically exhibits the same contraction mechanism.

Empirically, the approach produces sparser, subsystem-faithful models and competitive or superior accuracy. On HAI 1.0 ICS data, the binarized model achieved ACC 0.987, AUC 0.982, F1 0.987, and Youden's J 0.974, with coefficients concentrated on the attacked turbine loop, reducing false positives while preserving detection power. Synthetic experiments show near-perfect classification and precise lag localization under severe multicollinearity; additional studies (ionosphere, UNICEF) illustrate the interpretability/accuracy trade-off when phenomena deviate from threshold-and-lag regimes. Together, these results support binarization as a practical route to interpretable sparsity in complex systems.

Computationally, the transformation is linear-time in the number of feature values and trivially parallelizable; end-to-end fitting (transform + solver) was often faster than modeling raw features, consistent with the simplified correlation geometry and sparser solutions observed across datasets. This efficiency complements the interpretability benefits of explicit ranges and lags for each selected feature, enabling operational use in monitoring and control.

Limitations. Our guarantees are stated in the zero threshold PoC regime; extending closed-form results to nonzero thresholds remains open. The downstream model is linear/logistic; where decision boundaries are strongly nonlinear (e.g., UNICEF), accuracy gains may not materialize even though sparsity and attribution improve. Finally, negative-correlation edge cases can complicate inverse-norm monotonicity in stylized constant-correlation models; complement columns mitigate this in practice.

Outlook. Future work includes (i) formalizing correlation bounds for nonzero threshold binarization, (ii) an optimized parallel library for large-scale deployments, (iii) extensions to multiclass outcomes and structured sparsity (e.g., grouped lags), and (iv) broader longitudinal applications (e.g., chemical oscillators, climate, ecological thresholds) to further probe the method's scope while preserving transparency and efficiency.

References

[1] Freijeiro-González, L., Febrero-Bande, M., & González-Manteiga, W. (2022). A critical review of lasso and its derivatives for variable selection under dependence among covariates. *International Statistical Review, 90*(1), 118–145.

[2] Zhao, P., & Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research, 7*, 2541–2563.

[3] Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, 33*(1), 1–22.

[4] Feller, W. (1991). *An introduction to probability theory and its applications* (Vol. 2). John Wiley & Sons.

[5] Hotelling, H. (2012). Relations between two sets of variates. In S. Kotz & N. L. Johnson (Eds.), Breakthroughs in statistics: Methodology and distribution (pp. 162–190). Springer. (Original work published 1936)

[6] Horn, R. A., & Johnson, C. R. (2012). Matrix analysis (2nd ed.). Cambridge University Press.

[7] Shin, H., Lee, W., Yun, J., & Kim, H. (2020). HAI 1.0: HIL-based augmented ICS security dataset. In *Proceedings of the 13th USENIX Workshop on Cyber Security Experimentation and Test (CSET).*

[8] Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68*(1), 49–67.

[9] Turlach, B. A., Venables, W. N., & Wright, S. J. (2005). Simultaneous variable selection. *Technometrics, 47*(3), 349–363.

[10] Meier, L., van de Geer, S., & Bühlmann, P. (2006). The group lasso for logistic regression. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70(1), 53–71.

[11] Yang, Y., & Zou, H. (2015). A fast unified algorithm for solving group lasso penalized learning problems. *Statistics and Computing, 25*(6), 1129–1141.

[12] Kim, Y., Kim, J., & Kim, Y. (2006). Blockwise sparse regression. *Statistica Sinica, 16*, 375–390.

[13] Tibshirani, R., & Suo, X. (2016). An ordered lasso and sparse time-lagged regression. *Technometrics*, *58*(4), 415-423.

[14] Zhu, H., Xu, W., Zhang, Z., Xu, Y., & Fan, J. (2021). A variable selection approach for highly correlated predictors in high-dimensional genomic data. *Bioinformatics, 37*(16), 2238–2244.

[15] Rejchel, M., & Bogdan, M. (2020). Rank-based lasso: Efficient methods for high-dimensional robust model selection. *Journal of Machine Learning Research, 21*(244), 1–47.

[16] Ladha, L., & Deepa, T. (2011). Feature selection methods and algorithms. *International Journal on Computer Science and Engineering, 3*(5), 1787–1797.

[17] Katrutsa, A., & Strijov, V. (2017). Comprehensive study of feature selection methods to solve multicollinearity problem according to evaluation criteria. *Expert Systems with Applications, 76*, 1–11.

[18] Boros, E., Hammer, P. L., Ibaraki, T., Kogan, A., Mayoraz, E., & Muchnik, I. (2000). An implementation of logical analysis of data. *IEEE Transactions on Knowledge and Data Engineering, 12*(2), 292–306.

[19] Boros, E., Hammer, P. L., Ibaraki, T., & Kogan, A. (1997). Logical analysis of numerical data. *Mathematical Programming, 79*(1–3), 163–190.

[20] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, *101*(476), 1418-1429.

[21] Orender, Jason, Mohammad Zubair, and Jiangwen Sun. "LASSO Logic Engine: harnessing the logic parsing capabilities of the LASSO algorithm for longitudinal feature learning." *2022 IEEE International Conference on Big Data (Big Data)*. IEEE, 2022.