mathematics

MDPI

Article

Variable Selection for High-Dimensional Longitudinal Data via Within-Cluster Resampling

Yue Ma and Xuejun Jiang *

Department of Statistics and Data Science, Southern University of Science and Technology, Shenzhen 518055, China; 12131248@mail.sustech.edu.cn
* Correspondence: jiangxj@sustech.edu.cn

Abstract: The phenomenon of informative cluster size (ICS) emerges when the number of repeated measurements is correlated with the outcome variable. In such scenarios, the prevailing generalized estimating equation (GEE) method often yields biased estimates due to nonignorable cluster size. This study proposes an integrated methodology that explicitly accounts for ICS and provides a robust solution to mitigate its effects. Our approach combines within-cluster resampling (WCR) with a penalized likelihood framework, ensuring consistent model selection and parameter estimation across resampled datasets. Additionally, we introduce a penalized mean regression method to aggregate the estimators from multiple resampled datasets, producing a final estimator that improves the true positive discovery rate while controlling false positives. The proposed penalized likelihood method via WCR (PL_{WCR}) is evaluated through extensive simulations and an application to yeast cell-cycle gene expression data. The results demonstrate its robustness and superior performance in high-dimensional longitudinal data analysis with ICS.

Keywords: generalized estimating equations; longitudinal data analysis; informative cluster size; within-cluster resampling; penalized likelihood

MSC: 62H12; 62J12



Academic Editor: MinJae Lee

Received: 10 March 2025
Revised: 9 April 2025
Accepted: 12 April 2025
Published: 15 April 2025

Citation: Ma, Y.; Jiang, X. Variable Selection for High-Dimensional Longitudinal Data via Within-Cluster Resampling. *Mathematics* **2025**, *13*, 1293. https://doi.org/10.3390/math13081293

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Longitudinal data are characterized by repeated and correlated measurements across multiple covariates. Such data are ubiquitous in disciplines such as biomedical research, econometrics, and social sciences [1]. In [2], the authors highlighted prominent applications in these domains and discussed specialized statistical methodologies for analyzing longitudinal data. Conventional approaches are designed primarily for independent observations, and often fail to account for the intrinsic dependencies in longitudinal settings, resulting in biased parameter estimates and compromised inferential validity. Thus, advancing robust techniques that explicitly model the correlation structures of longitudinal data remains a critical methodological priority.

The generalized estimating equation (GEE) method [3] represents a cornerstone approach for longitudinal data analysis, offering consistent estimation through the incorporation of a working correlation structure. While this framework provides robustness against misspecification of the within-cluster correlation matrix, its validity critically depends on the mechanism used for missing data. Specifically, the GEE estimator may exhibit substantial bias when observations are missing not at random or missing at random, rather than missing completely at random. This limitation underscores the necessity of developing more sophisticated methodologies capable of addressing non-random missingness in

longitudinal studies. In longitudinal data analysis, informative cluster size (ICS) occurs when the conditional expectation of the response variable depends on the cluster size [4], a phenomenon also termed nonignorable cluster size. This issue arises frequently in applied settings. For example, the number of patient follow-up visits in clinical studies may correlate with disease severity. Formally, we can let X denote the covariates, Y represent the response variable (e.g., disease status), and M indicate the cluster size (e.g., number of visits). The ICS condition be formulated as $E(Y|X, M) \neq E(Y|X)$, which implies that standard analyses ignoring M may yield biased inferences. Similarly, observation frequency at sampling sites in ecological studies may depend on environmental factors. Another example arises in toxicology experiments where pregnant dams are randomly exposed to toxicants. Sensitive dams may produce litters with a higher rate of birth defects and experience more fetal resorptions, resulting in smaller litter sizes.

To address these challenges, Ref. [4] proposed the within-cluster resampling (WCR) method for marginal analysis of longitudinal data. This approach involves randomly selecting a single observation per cluster (with replacement) and applying standard inference techniques for independent data. The WCR method offers two key advantages over conventional GEE method. First, it is robust against correlation misspecification. By analyzing resampled independent observations, WCR inherently accommodates within-cluster correlations, thereby avoiding biases arising from misspecified working correlation structures. In contrast, the GEE method can produce unreliable estimates when the assumed correlation matrix is incorrect [5] or when the true dependence structure is complex or unknown. Second, WCR maintains validity in the presence of nonignorable cluster size. Unlike GEE, which implicitly weights observations by cluster size, WCR eliminates this weighting scheme by treating each cluster equally through resampling. This property ensures consistent estimation even when cluster sizes are informative about the response variable, as WCR's resampling mechanism inherently adjusts for ICS. Thus, WCR provides a principled framework for valid inference in settings where traditional GEE may fail due to either correlation misspecification or nonignorable cluster sizes.

Recent methodological developments have significantly advanced the analysis of high-dimensional longitudinal data, particularly in addressing the dual challenges of variable selection and parameter estimation. In [6], the authors introduced a Bayesian information criterion (BIC)-based selection approach using quadratic inference functions, while Ref. [7] developed corresponding asymptotic theory for binary outcomes when the number of predictors grows with cluster size. Building on this foundation, Ref. [8] proposed penalized generalized estimating equations (PGEE) with the smoothly clipped absolute deviation (SCAD) penalty [9], enabling simultaneous variable selection and estimation under specified moment conditions and working correlation structures. Parallel innovations include [10]'s quadratic decorrelated inference function for high-dimensional inference and [11]'s one-step debiased estimator via projected estimating equations, which facilitates hypothesis testing for linear combinations of high-dimensional coefficients. These contributions collectively provide a robust statistical framework for analyzing complex longitudinal data, helping to bridge critical gaps between theoretical development and practical application.

However, these GEE-based approaches critically depend on accurate estimation of within-cluster correlations, rendering them vulnerable to performance inconsistencies when informative cluster size (ICS) is present. To address this limitation, researchers have developed estimation techniques within the WCR framework. For instance, Ref. [12] introduced a modified WCR approach aimed at improving estimation efficiency, though requiring clusters of size greater than one and imposing further constraints on the correlation structure. Additionally, Ref. [13] developed a resampling-based cluster information criterion

for finite-dimensional semiparametric marginal mean regression. A naive yet essential strategy is to generate the final estimation by averaging a substantial quantity of resampled estimators. Specifically, when R resamplings are conducted, the WCR estimator is calculated as $R^{-1}\sum_{r=1}^{R}\hat{\boldsymbol{\beta}}^{(r)}$, where $\hat{\boldsymbol{\beta}}^{(r)}$ is obtained from the rth resampled data set for $r = 1,\ldots,R$. Nevertheless, this averaging strategy presents significant theoretical limitations in high-dimensional settings. On one hand, the method provides no formal guarantees for consistent model selection in high-dimensional regression problems; On the other, the unregularized averaging process may retain noise variables, potentially leading to model overfitting [14].

In this study, we introduce a novel model selection procedure tailored for high-dimensional longitudinal data in the presence of ICS. The dimension of covariates, denoted as p_n, is permitted to grow exponentially with the number of clusters n. Leveraging the stability that the WCR framework provides in finite-dimensional settings under ICS, we aim to extend its applicability to scenarios where the number of covariates surpasses the total number of clusters while ensuring both model selection and estimation consistency. Our proposed penalized likelihood via WCR approach (abbreviated as PL_{WCR}) combines penalized likelihood with WCR through three key steps. For resampling, following [4], we draw one observation per cluster (with replacement), creating n independent samples. For regularized estimation, we perform simultaneous variable selection and parameter estimation for each of R resamples via penalized likelihood maximization, utilizing the resulting n independent observations of resampling data sets. Finally, for stable aggregation, we prevent overfitting from naive averaging by applying component-wise penalized mean regression across the R estimates, producing a sparse final model. This integrated framework maintains WCR's robustness to ICS while addressing high-dimensional challenges through proper regularization.

Our PL_{WCR} approach offers three distinct improvements over existing longitudinal data approaches. (i) Correlation-robust inference: PL_{WCR} naturally handles intracluster dependence without explicit correlation modeling while avoiding estimation bias from misspecified correlation structures, This is especially crucial in higher dimensions. (ii) ICS-robust estimation: Our method maintains robustness even when dependencies exist between the response variables and cluster sizes, effectively addressing the challenges associated with ICS. (iii) Stable sparse recovery: By incorporating penalized mean regression during aggregation, PL_{WCR} ensures sparse solutions and significantly reduces overfitting risk versus naive averaging These advantages collectively enhance the reliability and applicability of the proposed method in complex longitudinal data analyses.

The rest of this paper is structured as follows. Section 2 presents the model framework, examines the limitations of GEE estimation, and details our methodology. Theoretical guarantees of the proposed estimator under mild regularity conditions are established in Section 3. Comprehensive simulation results appear in Section 4, followed by an application to yeast cell-cycle gene expression data in Section 5. We conclude with discussion of broader implications in Section 6, while technical proofs are collected in the Appendix A.

Below, we introduce the notation commonly used throughout this paper. For any constant a, its integer part is denoted by $[a]$. For any vector $\mathbf{a} = (a_1,\ldots,a_m)^T \in \mathbb{R}^m$ and $\mathbf{b} = (b_1,\ldots,b_m)^T \in \mathbb{R}^m$, we define $|\mathbf{a}| = (|a_1|,\ldots,|a_m|)^T$, $\|\mathbf{a}\|_1 = \sum_{i=1}^{m}|a_i|$, $\|\mathbf{a}\|_2 = \sqrt{\sum_{i=1}^{m}a_i^2}$, $|\mathbf{a}|_\infty = \max_{1\le i\le m}|a_i|$, and $\mathbf{a}\circ\mathbf{b} = (a_1b_1,\ldots,a_m\,b_m)^T$. For any subset $\mathcal{M}$ of the row index set $\{1,\ldots,m\}$ of $\mathbf{a}$, $|\mathcal{M}|$ denotes the cardinality of $\mathcal{M}$ and $\mathbf{a}_{\mathcal{M}}$ represents the subvector of $\mathbf{a}$ corresponding to the indices in $\mathcal{M}$, with the indicator function denoted by $I(\cdot)$. Let $\mathbf{B}$ be an $m\times m$ matrix and let $\mathbf{B}^{-1}$ be its inverse. For subsets $\mathcal{M}_1,\mathcal{M}_2 \subseteq \{1,\ldots,m\}$, $\mathbf{B}_{\mathcal{M}_1}$ denotes the submatrix of $\mathbf{B}$ formed by column indices in $\mathcal{M}_1$ and $\mathbf{B}_{\mathcal{M}_1,\mathcal{M}_2}$ represents the submatrix with rows and columns indexed by $\mathcal{M}_1$ and $\mathcal{M}_2$, respectively. The maximum

and minimum eigenvalues of a matrix are denoted by $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$, respectively, while the spectral norm is provided by $\|\mathbf{B}\|_2 = \sqrt{\lambda_{\max}(\mathbf{B}^T\mathbf{B})}$.

2. Methodology

2.1. Problem Formulation

Let Y_{ij} represent the response variable of the jth observation in the ith cluster, with a p_n-dimensional covariate vector $\mathbf{X}_{ij} = (X_{ij,1}, \ldots, X_{ij,p_n})^T$, where $i = 1, \ldots, n$, $j = 1, \ldots, M_i$, and M_i refers to cluster size. Given $\mathbf{X}_{ij}$, we assume that the probability density function of Y_{ij} has the following expression:

$$f(Y_{ij}|\mathbf{X}_{ij}, \boldsymbol{\beta}^*, \phi) = \exp\{[Y_{ij}\mathbf{X}_{ij}^T\boldsymbol{\beta}^* - b(\mathbf{X}_{ij}^T\boldsymbol{\beta}^*)]/\phi + c(Y_{ij}, \phi)\} \tag{1}$$

where $\boldsymbol{\beta}^*$ is an unknown $\mathbb{R}^{p_n}$ coefficient vector, $b(\cdot)$ is a known link function, and ϕ is a dispersion parameter. Observations within each cluster are commonly assumed to be correlated, but observations are independent across clusters. Here, we allow the dimensionality p_n to be larger than the number of culsters n, i.e., $\log(p_n) = O(n^\alpha)$ for $0 < \alpha < 1$; the true population parameter $\boldsymbol{\beta}^*$ is sparse, with only several nonzero components. We define $\mathbf{b}_i'(\boldsymbol{\beta}) = (b'(\mathbf{X}_{i1}^T\boldsymbol{\beta}), \ldots, b'(\mathbf{X}_{iM_i}^T\boldsymbol{\beta}))^T$ and $\mathbf{b}_i''(\boldsymbol{\beta}) = \mathrm{diag}(b''(\mathbf{X}_{i1}^T\boldsymbol{\beta}), \ldots, b''(\mathbf{X}_{iM_i}^T\boldsymbol{\beta}))$. Let $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iM_i})^T$ and $\mathbf{X}_i = (\mathbf{X}_{i1}, \mathbf{X}_{i2}, \ldots, \mathbf{X}_{iM_i})^T$.

To obtain the optimal estimation, [3] proposed the GEE:

$$n^{-1}\sum_{i=1}^{n}\{\mathbf{X}_i^T\{\mathbf{b}_i''(\boldsymbol{\beta})\}^{1/2}\mathbf{V}_i^{-1}\{\mathbf{b}_i''(\boldsymbol{\beta})\}^{-1/2}(\mathbf{Y}_i - \mathbf{b}_i'(\boldsymbol{\beta}))\} = \mathbf{0} \tag{2}$$

where $\mathbf{V}_i$ is the covariance matrix of $\mathbf{Y}_i$. A naive choice of the covariance matrix is the working independent structure, i.e., $\mathbf{V}_i = \mathbf{I}_{M_i}$. However, the validness of the GEE is built upon the assumption that M_i is independent of $\{Y_{ij}, \mathbf{X}_{ij}, i = 1, \ldots, n, j = 1, \ldots, M_i\}$. This assumption is violated when the cluster size is informative [4], i.e., $E(Y_{ij} \mid \mathbf{X}_{ij}, M_i) \neq E(Y_{ij} \mid \mathbf{X}_{ij})$. With the presence of ICS, it follows that

$$E\{\mathbf{X}_i^T\{\mathbf{b}_i''(\boldsymbol{\beta}^*)\}^{1/2}\mathbf{V}_i^{-1}\{\mathbf{b}_i''(\boldsymbol{\beta}^*)\}^{-1/2}(\mathbf{Y}_i - \mathbf{b}_i'(\boldsymbol{\beta}^*))\} = E\{M_iE(\mathbf{X}_{i1}\varepsilon_{i1} \mid M_i)\} \neq \mathbf{0}, \tag{3}$$

and by the law of iterated expectation, $\varepsilon_{i1} = Y_{i1} - b'(\mathbf{X}_{i1}^T\boldsymbol{\beta}^*)$. From (3), the GEE leads to a biased estimation. This motivates us to propose a novel estimation method for longitudinal data when the cluster size is informative.

2.2. Penalized WCR Likelihood Estimation

For the model in (1), the scale-shift likelihood function of $\{Y_{ij}, \mathbf{X}_{ij}, i = 1, \ldots, n, j = 1, \ldots, M_i\}$ is

$$L_n(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{M_i} Y_{ij}\mathbf{X}_{ij}^T\boldsymbol{\beta} - b(\mathbf{X}_{ij}^T\boldsymbol{\beta}). \tag{4}$$

By taking the derivative, we obtain

$$\partial L_n(\boldsymbol{\beta})/\partial\boldsymbol{\beta} = n^{-1}\sum_{i=1}^{n}\{\mathbf{X}_i^T\{\mathbf{b}_i''(\boldsymbol{\beta})\}^{1/2}\mathbf{V}_i^{-1}\{\mathbf{b}_i''(\boldsymbol{\beta})\}^{-1/2}(\mathbf{Y}_i - \mathbf{b}_i'(\boldsymbol{\beta}))\},$$

where $\mathbf{V}_i = \mathbf{I}_{M_i}$ is the assumed working independence matrix. If we let $\partial L_n(\boldsymbol{\beta})/\partial\boldsymbol{\beta} = \mathbf{0}$, it is exactly the GEE (2). This demonstrates that the naive likelihood approach also enjoys the excellent properties of the GEE. However, this does not happen when ICS is present, as demonstrated by (3). The effect of ICS originates from the randomness of $\sum_{j=1}^{M_i} Y_{ij}\mathbf{X}_{ij}^T\boldsymbol{\beta} - b(\mathbf{X}_{ij}^T\boldsymbol{\beta})$ in (4). To eliminate the effect of ICS, we adopt the WCR idea proposed by [4].

Let Z_i be a random variable independent of $\{Y_{ij}, \mathbf{X}_{ij}, i = 1, \ldots, n, j = 1, \ldots, M_i\}$ and let z_i be a sample of Z_i which takes values in $\{1, \ldots, M_i\}$ with probability $P(z_i = j) = 1/M_i$, $j = 1, \ldots, M_i$. WCR randomly draws an observation from each cluster with replacement; thus, the likelihood function via WCR is

$$\tilde{L}_n(\boldsymbol{\beta}; \mathbf{z}) = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{M_i} I(j = z_i) \{ Y_{ij} \mathbf{X}_{ij}^T \boldsymbol{\beta} - b(\mathbf{X}_{ij}^T \boldsymbol{\beta}) \}, \tag{5}$$

where $I(\cdot)$ is the indicator function and $\mathbf{z} = (z_1, \ldots, z_n)$. With straightforward calculations and using the law of iterative expectation, we can obtain

$$E\{\partial \tilde{L}_n(\boldsymbol{\beta}^*; \mathbf{z}) / \partial \boldsymbol{\beta}\} = E\left(\mathbf{X}_{iz_i} \varepsilon_{iz_i}\right) = \mathbf{0}.$$

This guarantees that the resulting maximum likelihood estimation is consistent. Benefiting from this property, we can develop variable selection and simultaneous estimation for the model in (1). The penalized likelihood function via WCR is defined as

$$Q_n(\boldsymbol{\beta}; \mathbf{z}) = \tilde{L}_n(\boldsymbol{\beta}; \mathbf{z}) - \sum_{d=1}^{p_n} p_{\lambda_n}(|\beta_d|), \tag{6}$$

where $p_{\lambda_n}(\cdot)$ is a penalty function indexed by a regularization parameter λ_n. We generate the R independent observations $\{\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(R)}\}$ of $\mathbf{z}$. Let $\tilde{\boldsymbol{\beta}}^r$ be the maximizer of $Q_n(\boldsymbol{\beta}, \mathbf{z}^{(r)})$, i.e.,

$$\tilde{\boldsymbol{\beta}}^{(r)} = \arg\max_{\boldsymbol{\beta}} Q_n(\boldsymbol{\beta}; \mathbf{z}^{(r)}) \tag{7}$$

for $r = 1, \ldots, R$. The next issue we need to focus on is how to integrate the R estimators. A naive approach is to average the R estimators, as proposed in [4]. However, this trivial averaging approach may cause many covariates to have nonzero estimated coefficients, resulting in model overfitting. Alternatively, we propose aggregating these R estimators by regularization. Therefore, we consider the following penalized least squares estimation:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \{ R^{-1} \sum_{r=1}^{R} \| \tilde{\boldsymbol{\beta}}^{(r)} - \boldsymbol{\beta} \|_2^2 + \sum_{d=1}^{p_n} \lambda_n^{(d)} |\beta_d| \}. \tag{8}$$

We summarize our penalized likelihood (PL) via WCR (PL_{WCR}) method in Algorithm 1.

Algorithm 1: PL_{WCR} Algorithm

Input: $\{(\mathbf{Y}_i, \mathbf{X}_i)\}_{i=1}^n$, λ_n and $\{\lambda_n^{(1)}, \ldots, \lambda_n^{(p_n)}\}$
Result: $\hat{\mathcal{S}}$ and $\hat{\boldsymbol{\beta}}$
Step 1: Penalized likelihood estimation via WCR
for $r = 1, \ldots, R$, **do**
  Randomly draw an observation $\mathbf{z}^{(r)}$;
  Compute the estimator $\tilde{\boldsymbol{\beta}}^{(r)}$ by solving
  $\tilde{\boldsymbol{\beta}}^{(r)} = \arg\max_{\boldsymbol{\beta}} \{ \tilde{L}_n(\boldsymbol{\beta}; \mathbf{z}) - \sum_{d=1}^{p_n} p_{\lambda_n}(|\beta_d|) \}$;
end
Step 2: Aggregation
On the basis of $\{\tilde{\boldsymbol{\beta}}^{(r)}\}_{r=1}^R$, calculate the aggregate estimates $\hat{\boldsymbol{\beta}}$ as
$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \{ R^{-1} \sum_{r=1}^{R} \| \tilde{\boldsymbol{\beta}}^{(r)} - \boldsymbol{\beta} \|_2^2 + \sum_{d=1}^{p_n} \lambda_n^{(d)} |\beta_d| \}$;
Then let $\hat{\mathcal{S}} = \{ d : |\hat{\beta}_d| > 0 \}$.

3. Theoretical Properties

In this section, we establish the consistency of model selection and parameter estimation.

For convenience, we first introduce some notation. Let $\mathcal{S}$ be the index set corresponding to the nonzero components in $\boldsymbol{\beta}^*$. We define $\varepsilon_{ij} = Y_{ij} - b'(\mathbf{X}_{i1}^T\boldsymbol{\beta}^*)$, $\mathcal{N}_0 = \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \leq \sqrt{s_n/n}, \boldsymbol{\beta}_{\mathcal{S}^c} = \mathbf{0}\}$, and $\mathcal{P} = \{\mathbf{z} = (z_1, \ldots, z_n) : z_i \in \{1, 2 \ldots, M_i\}, i = 1, \ldots, n\}$. We denote $\boldsymbol{\Sigma}_1(\boldsymbol{\beta}; \mathbf{z}) = n^{-1}\sum_{i=1}^n b''(\mathbf{X}_{i1}^T\boldsymbol{\beta})\mathbf{X}_{iz_i,\mathcal{S}}\mathbf{X}_{iz_i,\mathcal{S}}^T$ and $\boldsymbol{\Sigma}_{21}(\boldsymbol{\beta}; \mathbf{z}) = n^{-1}\sum_{i=1}^n b''(\mathbf{X}_{i1}^T\boldsymbol{\beta})\mathbf{X}_{iz_i,\mathcal{S}^c}\mathbf{X}_{iz_i,\mathcal{S}}^T$. Let $s_n = |\mathcal{S}|$ and $d_n = \min_{d\in\mathcal{S}}|\beta_d^*|/2$.

The following regularity conditions are required for establishing asymptotic results.

(A1) Let $\rho_{\lambda_n}(t) = \lambda_n^{-1}p_{\lambda_n}(t)$ for $\lambda_n > 0$. Assume that $\rho_{\lambda_n}(t)$ is increasing and concave in $t \in [0, \infty)$ and that it has a continuous derivative $\rho'_{\lambda_n}(t)$ with $\rho'_{\lambda_n}(0+) > 0$. In addition, $\rho'_{\lambda_n}(0+)$ is independent of λ_n; for simplicity, we write it as $\rho'(0+)$.

(A2) (i) $\max_{1\leq i\leq n}\max_{1\leq j\leq M_i} E\{\exp(|\varepsilon_{ij}|/c_1) - 1 - |\varepsilon_{ij}|/c_1 \mid \mathbf{X}_{ij}\}c_1^2 \leq \frac{c_2}{2}$ for some positive constants c_1 and c_2.
(ii) $\|\mathbf{a}^T\mathbf{X}_{ij}\|_{\psi_2} \leq c_6\|\mathbf{a}\|_2$ for any $\mathbf{a} \in \mathbb{R}^{p_n}$, where $\|\mathbf{a}^T\mathbf{X}_{ij}\|_{\psi_2} = \inf_{t>0}\{E\exp(t^{-2}|\mathbf{a}^T\mathbf{X}_{ij}|^2) \leq 2\}$.

(A3) Assume that for any $\mathbf{z} \in \mathcal{P}$ there exist some positive constants c_3 and c_4 such that $\min_{\boldsymbol{\beta}\in\mathcal{N}_0}\lambda_{\min}(\boldsymbol{\Sigma}_1(\boldsymbol{\beta}; \mathbf{z})) \geq c_3$ and $\max_{\boldsymbol{\beta}\in\mathcal{N}_0}\sup_{\mathbf{a}\in\mathbb{R}^{s_n}, \|\mathbf{a}\|_2\leq 1}\|\boldsymbol{\Sigma}_{21}(\boldsymbol{\beta}; \mathbf{z})\mathbf{a}\|_\infty \leq c_4$ holds with probability going to one.

(A4) Assume that $\lambda_n = o(d_n)$, $\max\{\sqrt{s_n/n}, \sqrt{\log(p_n)/n}\} = o(\lambda_n)$ and $p'_{\lambda_n}(d_n) = O(n^{-1/2})$.

(A5) Define the supremum of the local concavity of ρ_{λ_n} as

$$\kappa(\rho_{\lambda_n}) = \sup_{\mathbf{b}\in\mathcal{N}_0}\lim_{a\to 0+}\max_{k\in\mathcal{S}}\sup_{t_1<t_2\in(|b_k|-a,|b_k|+a)} -\frac{\rho'_{\lambda_n}(t_2) - \rho'_{\lambda_n}(t_1)}{t_2 - t_1}$$

and suppose that $\lambda_n\kappa(\rho_{\lambda_n}) = o(1)$.

Condition (A1) is satisfied by commonly used folded concave penalty functions such as LASSO [15], SCAD [9], and MCP [16] (with $a \geq 1$). In [9], the authors state that an ideal estimator should be equipped with three desirable properties: unbiasedness, sparsity, and continuity. However, the LASSO penalty fails to reach an unbiased estimator, the L_q penalty with $q > 1$ does not produce a sparse solution, and the L_q penalty with $0 \leq q < 1$ does not satisfy the continuous condition. In our study, SCAD penalty is chosen to conduct variable selection. Its derivative is defined as

$$p'_{\lambda_n}(t) = \lambda_n\left\{I(t \leq \lambda_n) + \frac{(a\lambda_n - t)_+}{(a-1)\lambda_n}I(t > \lambda_n)\right\}$$

for some $a > 2$ and $t \geq 0$. This penalty satisfies both Condition (A1) and the three desirable properties within the penalized likelihood framework.

Condition (A2) restricts the moment of errors and the sub-Gaussian distribution of the predictors to satisfy this condition, which is commonly used for high-dimensional penalized regression [17–19]. Condition (A2)(i) holds for the exponential generalized linear model family (1).

In Condition (A3), the singular value constraint $\min_{\boldsymbol{\beta}\in\mathcal{N}_0}\lambda_{\min}(\boldsymbol{\Sigma}_1(\boldsymbol{\beta}, \mathbf{z})) \geq c_1$ and $\max_{\boldsymbol{\beta}\in\mathcal{N}_0}\sup_{\mathbf{a}\in\mathbb{R}^{s_n}, \|\mathbf{a}\|_2\leq 1}\|\boldsymbol{\Sigma}_{21}(\boldsymbol{\beta}, \mathbf{z})\mathbf{a}\|_\infty \leq c_4$ is the same as (26) and (27) in [17], which has been proven for the family of exponential generalized linear models in [18] when the covariate vector follows from a sub-Gaussian distribution. Furthermore, $\|\boldsymbol{\Sigma}_{21}(\boldsymbol{\beta}, \mathbf{z})\mathbf{a}\|_\infty \leq c_4$ is the unrepresentable condition [20], which is commonly assumed for penalized regression.

Condition (A4) restricts the strength of signals, number of nonzero covariates, dimension of the feature space, and tuning parameter, which is commonly used in [17,18].

From Condition (A4), we allow $\log(p_n) = O(n^\alpha)$, $0 < \alpha < 1$, which is a common condition for ultra-high-dimensional feature spaces in the literature. Finally, Condition (A5) is assumed by [17] in order to ensure the second-order condition.

The following theorem establishes the oracle property of the penalized likelihood estimator $\tilde{\boldsymbol{\beta}}^{(r)}$ via WCR.

Theorem 1. *Assume that Conditions (A1)–(A5) hold. Then, we have* $\|\tilde{\boldsymbol{\beta}}^{(r)} - \boldsymbol{\beta}^*\|_2 = O_P(\sqrt{s_n/n})$ *and* $P(\hat{\mathcal{S}}^{(r)} = \mathcal{S}) \to 1$ *as* $n \to \infty$, *where* $\hat{\mathcal{S}}^{(r)} = \{d : \tilde{\beta}_d^{(r)} \neq 0, d = 1, \ldots, p_n\}$.

Theorem 1 establishes model selection sparsity and the estimation consistency of $\tilde{\boldsymbol{\beta}}^{(r)}$. It also holds for the ultra-high-dimensional feature space regardless of whether or not the cluster size is informative.

In the following, we establish the oracle property of the penalized likelihood estimator $\hat{\boldsymbol{\beta}}$ via WCR.

Theorem 2. *Assume that the conditions in Theorem 1 all hold. If* $\lambda_n^{(d)}$ *satisfies* $\lambda_n^{(d)} = o(d_n)$ *and* $\sqrt{s_n/n} = o(\lambda_n^{(d)})$, *then we have* $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 = O_P(\sqrt{s_n/n})$ *and* $P(\hat{\mathcal{S}} = \mathcal{S}) \to 1$ *as* $n \to \infty$, *where* $\hat{\mathcal{S}} = \{d : \hat{\beta}_d \neq 0, d = 1, \ldots, p_n\}$.

Compared to the penalized GEE method proposed by [8], our PL_{WCR} method demonstrates several significant advantages in handling high-dimensional longitudinal data with informative cluster sizes.

First, PL_{WCR} substantially improves model selection accuracy, particularly in the presence of ICS. By adopting the WCR framework described in [4], our method assigns equal weights to observations from clusters of varying sizes, thereby ensuring robustness when the cluster sizes are related to outcomes. Repeated marginal analyses conducted on resampled datasets allow us to fully utilize sample information while maintaining consistent screening results. The subsequent aggregation of R candidate models through component-wise penalized least squares further enhances the reliability of our approach. Importantly, this aggregation process ensures that occasional omissions of important covariates in individual analyses do not substantially affect the final estimates, while repeated misidentification of noise variables is effectively suppressed. As a result, PL_{WCR} achieves a higher true positive rate while simultaneously reducing false positives. This represents a distinct advantage over the PGEE method, which relies on a single model fitting that provides each variable with only one selection opportunity, and is consequently more vulnerable to selection errors.

Second, PL_{WCR} offers superior handling of within-cluster correlations without requiring explicit specification of the correlation structure. Though computationally intensive, the WCR approach effectively circumvents the need to estimate potentially complex intra-cluster correlation structures. This represents a significant improvement over the PGEE method, where accurate estimation critically depends on within-cluster correlations. This requirement often proves problematic, particularly in nonlinear cases where correlation estimates may be invalid. The implicit treatment of correlation in PL_{WCR} not only simplifies implementation but also enhances the method's robustness across diverse data scenarios. These combined advantages position PL_{WCR} as a more reliable and versatile tool for high-dimensional longitudinal data analysis compared to existing penalized GEE approaches.

4. Simulation Studies

In this section, we describe a series of numerical experiments evaluating the performance of the proposed PL_{WCR} method in model selection for longitudinal data. The tuning

parameter within the penalized likelihood is selected by minimizing the extended Bayesian information criterion (BIC) [21], ensuring an optimal balance between model complexity and goodness of fit. In the aggregation step, we investigate several different choices of resampling time ($R = 50, 100, 200, 500, 1000$), providing sensitivity analyses to demonstrate the robustness and tradeoffs between computational efficiency and stability, on the basis of which we choose $R = 500$ times resampling for the most robust and reliable results. Furthermore, with the exception of LASSO, we explore the SCAD penalty as an alternative sparse penalty function in order to elucidate potential improvements or limitations. We use five-fold cross-validation to select $\lambda_n^{(d)}$ in (8) for each $d = 1, \ldots, p_n$. In addition, we compare the aggregation performance of the simple averaging method. Specifically, we take the average of the resulting R estimated values for each covariate component. These experiments aimed to validate the effectiveness and stability of our $\mathrm{PL}_{\mathrm{WCR}}$ approach in handling high-dimensional longitudinal data with informative cluster sizes.

In this study, we compare our proposed method with two competing approaches. The first is the penalized GEE (PGEE) method introduced by [8], which is implemented using the R package `PGEE`. Three commonly used working correlation structures are considered: independence, exchangeable (equally correlated), and autoregressive of order 1 (AR-1). To facilitate a clear comparison of these methods under different correlation scenarios, we append the suffixes ".indep", ".exch", and ".ar1" to denote the independence, exchangeable, and AR-1 correlation structures, respectively. As recommended by [8], the PGEE method is executed over 30 iterations and a fourfold cross-validation procedure is employed to select the tuning parameter in the SCAD penalty function. Additionally, following [8], a coefficient is identified as zero if its estimated magnitude falls below the threshold value of 10^{-3}. This comparative framework ensures a rigorous evaluation of the proposed method's performance relative to established approaches. Additionally, we consider a simplified approach that neglects within-cluster correlation and applies a penalized maximum likelihood with the L_1-penalty, referred to as naive LASSO. This method serves as a baseline for comparison, particularly in scenarios where intracluster dependencies are ignored. Our investigations encompass both correlated continuous and binary response variables.

Notably, the penalized generalized estimating equations (PGEE) method requires approximately 4 h to fit a single generated dataset, even when the covariate dimension p_n is as low as 50. In contrast, our proposed $\mathrm{PL}_{\mathrm{WCR}}$ method completes the same task in just 5 min, demonstrating significantly improved computational efficiency. Due to the substantial time demands of PGEE, we exclude its performance evaluation for datasets with $p_n = 500$. This omission highlights the practical advantages of our $\mathrm{PL}_{\mathrm{WCR}}$ approach in high-dimensional settings where computational efficiency is critical.

We generate 100 longitudinal datasets for each setup. To evaluate the screening performance, we calculate the following criteria: (1) true positive (TP), the average number of true variables that are correctly identified; (2) false positive (FP), the average number of unimportant variables that are selected by mistake; (3) coverage rate (CR), the probability that the selected model covers the true model; and (4) mean square error (MSE), calculated as $\sum_{r=1}^{100} \|\widehat{\boldsymbol{\beta}}^r - \boldsymbol{\beta}\|^2/100$, where $\widehat{\boldsymbol{\beta}}^r$ is the estimate obtained on the rth generated dataset. All algorithms were run on a computer with an Intel(R) Xeon(R) Gold 6142 CPU and 256 GB RAM.

Example 1. *In this example, we consider the underlying model as*

$$Y_{ij} = U_{ij}\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \epsilon_{ij}$$

for correlated normal responses with ICS, where $i = 1, \ldots, 200$ *and* $j = 1, \ldots, M_i$. *The random cluster size* M_i *takes a value in the set* $\{2, 4, 15\}$, *with the probability distribution as* $P(M_i = 2) = 9/16$, $P(M_i = 4) = 3/8$, *and* $P(M_i = 15) = 1/16$. *Here, we define* $U_{ij} = 1(M_i \leq 4)$. *The* p_n*-dimensional vector of covariates* $\boldsymbol{X}_{ij} = (x_{ij,1}, \ldots, x_{ij,p_n})^T$ *are independently generated from a multivariate normal distribution, which has mean 0 and an autoregressive covariance matrix with marginal variance 1 and autocorrelation* 0.4. *The coefficient vector* $\boldsymbol{\beta} = (2, -1, 1.5, -2, 0, 0, \ldots, 0)^T$ *presents a homogeneous effect. The random errors* $(\epsilon_{i1}, \ldots, \epsilon_{iM_i})^T$ *obey a multivariate normal distribution with marginal mean 0, marginal variance 1, and an exchangeable correlation matrix with parameter* $\rho = 0.5$ or 0.8.

In Example 1, we control the ICS severity by introducing a univariable U_{ij}, leading to

$$E(Y_{ij}|\mathbf{X}_{ij}, M_i) = U_{ij}\mathbf{X}_{ij}^T\boldsymbol{\beta},$$

which is not equal to

$$E(Y_{ij}|\mathbf{X}_{ij}) = E(U_{ij})\mathbf{X}_{ij}^T\boldsymbol{\beta}.$$

Since the distribution of U_{ij} depends on the cluster size M_i, the marginal expectation of the responses is also influenced by M_i. By adjusting the distributions of M_i and U_{ij}, the severity of informative cluster size (ICS) can be modified.

Table 1 presents a comparative summary of model selection performance and estimation accuracy for the three competing methods in Example 1. While all methods successfully identify the TPs, only our PL_{WCR} method consistently selects a model that closely approximates the true sparse structure. For $p_n = 50$, the PGEE method incorporates a small number of redundant variables under the three commonly used working correlation structures. Similarly, the naive LASSO approach selects a moderately sized model, particularly when the within-cluster correlation increases to 0.8. Notably, the PL_{WCR} method achieves the smallest mean squared error (MSE) while demonstrating robustness against increases in both model dimension and intracluster correlation. In contrast, PGEE exhibits significant inconsistency in the presence of ICS, with estimation errors nearly doubling as the within-cluster correlation intensifies. Furthermore, naive LASSO produces biased estimates and displays sensitivity to the dimension of the covariates, resulting in a substantially larger MSE. These findings underscore the superior performance of our PL_{WCR} method in high-dimensional longitudinal data analysis under ICS.

Table 2 provides model selection and parameter estimation results at $\rho = 0.5$ and $p_n = 500$ for Example 1 with varying resampling times, two alternative aggregation penalties, and the simple averaging method. For linear regression, LASSO exhibits decreasing FPs and relatively robust MSE with increased resampling times. SCAD retains zero FPs and equally efficient estimates to LASSO. When compared with conventional L_1 regularization, the SCAD penalty possesses a superior sparsity-inducing property. However, simple averaging aggregation yields non-sparse models. With increasing sampling frequency, a growing number of redundant variables with weak signals are identified.

Example 2. *In this example, we consider correlated binary responses with marginal mean satisfying*

$$E(Y_{ij} \mid \boldsymbol{X}_{ij}, M_i) = U_{ij} \exp(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}) / \{1 + \exp(\boldsymbol{X}_{ij}^T\boldsymbol{\beta} + \log(15/16))\},$$

where $i = 1, \ldots, 200$ *and* $j = 1, \ldots, M_i$. *To account for ICS, we let the random cluster size* M_i *follow the probability distribution as* $P(M_i = 4) = 9/16$, $P(M_i = 6) = 3/8$, *and* $P(M_i = 10) = 1/16$. *The components of the covariates vector* $\boldsymbol{X}_{ij} = (x_{ij,1}, \ldots, x_{ij,p_n})^T$ *are jointly normal distributed with mean 0 and an autoregressive covariance matrix. The covariance matrix has marginal variance 1 and autocorrelation 0.4. We set* $\boldsymbol{\beta} = (1, -0.9, 0.7, 0, 0, \ldots, 0)^T$. *The corre-*

lated binary responses were generated using the R package `SimCorMultRes` *using an exchangeable within-cluster correlation parameter* $\rho = 0.5$ or 0.8.

Table 1. Model selection results for Example 1 with ICS.

p_n	ρ	Approach	TP	FP	CR	MSE
50	0.5	PL_{WCR}	4.00 (0.00)	0.85 (1.08)	1.00 (0.00)	0.033 (0.022)
		PGEE.indep	4.00 (0.00)	7.13 (3.92)	1.00 (0.00)	0.395 (0.117)
		PGEE.exch	4.00 (0.00)	14.79 (9.35)	1.00 (0.00)	0.596 (0.287)
		PGEE.ar1	4.00 (0.00)	16.29 (8.40)	1.00 (0.00)	0.513 (0.143)
		naive lasso	4.00 (0.00)	2.55 (1.95)	1.00 (0.00)	0.979 (0.215)
	0.8	PL_{WCR}	4.00 (0.00)	0.70 (0.88)	1.00 (0.00)	0.032 (0.021)
		PGEE.indep	4.00 (0.00)	6.69 (3.18)	1.00 (0.00)	0.406 (0.109)
		PGEE.exch	4.00 (0.00)	19.44 (13.51)	1.00 (0.00)	1.099 (1.139)
		PGEE.ar1	4.00 (0.00)	21.31 (14.05)	1.00 (0.00)	1.166 (1.066)
		naive lasso	4.00 (0.00)	2.69 (2.41)	1.00 (0.00)	0.978 (0.225)
500	0.5	PL_{WCR}	4.00 (0.00)	0.19 (0.44)	1.00 (0.00)	0.032 (0.018)
		naive lasso	4.00 (0.00)	2.93 (2.01)	1.00 (0.00)	1.430 (0.278)
	0.8	PL_{WCR}	4.00 (0.00)	0.27 (0.57)	1.00 (0.00)	0.032 (0.019)
		naive lasso	4.00 (0.00)	3.09 (2.19)	1.00 (0.00)	1.426 (0.295)

Correlated binary outcomes inherently contain less information than continuous data, posing greater challenges when seeking to accurately identify significant variables and obtain precise estimates. The model selection and estimation results for Example 2 are summarized in Table 3. Notably, the proposed PL_{WCR} method consistently identifies all relevant features while maintaining model sparsity across all scenarios. In contrast, the PGEE method struggles to identify the true model under an independent correlation structure, with the CR rate dropping below 60%. Although PGEE achieves asymptotic consistency under the other two intracluster correlation structures, this comes at the expense of higher FPs compared to our method. Furthermore, naive LASSO exhibits some deficiencies in identifying TPs when the covariates are autoregressively correlated with a coefficient of 0.4. In the presence of ICS, the MSE of our PL_{WCR} method experiences minor adverse effects but remains the lowest among the three competing methods. For $p_n = 50$, PGEE produces severely biased estimates when the observations are assumed to be independent. Even when the working correlation structure is correctly specified, the MSE of PGEE is approximately three times higher than that of PL_{WCR}. Meanwhile, naive LASSO demonstrates significant bias and yields the largest MSE, highlighting its limitations in handling correlated binary outcomes under ICS. These results underscore the robustness and precision of our PL_{WCR} method in high-dimensional longitudinal data analysis with binary responses.

Sensitivity analyses of Example 2 at $\rho = 0.5$ and $p_n = 500$ are presented in Table 4. Obviously, LASSO shows slight FPs in contrast to SCAD's constant zero FPs. For binary outcomes, the SCAD penalty demonstrates negligible estimation deviations compared with LASSO at $R = 50, 100, 200$. The estimation errors plateau as R increases, indicating that the aggregation effect becomes independent of the penalty function when the resampling times reach 500. This demonstrates that our proposed PL_{WCR} method exhibits asymptotic robustness with respect to the number of resamplings. Unlike binary models, simple averaging aggregation generates a substantial number of redundant variables, which severely compromises estimation efficiency and doubles the MSE of the estimates.

The performance of the penalized GEE approach varies under different working correlation structures (independence, exchangeable, and autoregressive), particularly in the presence of informative cluster size (ICS). Even when the working correlation structure is correctly specified, the resulting estimates can still show substantial bias, as illustrated in

Tables 1 and 3. The naive LASSO method is also affected by both intracluster correlations and ICS. In practice, the true within-cluster correlation structure is often complex and challenging to identify. To address these issues, our proposed PL_{WCR} method leverages within-cluster resampling to simultaneously account for intricate correlation patterns and ICS, leading to more robust model selection and parameter estimation. Moreover, the proposed PL_{WCR} method maintains its robustness across different levels of dimensionality from $p_n = 50$ to 500 thanks to the stabilizing effect of aggregation. In contrast, the penalized GEE fails to produce a sparse model, while naive LASSO suffers from considerable estimation bias even at moderate dimensionality $p_n = 50$.

Table 2. Sensitivity analyses for PL_{WCR} with alternative aggregation for Example 1 with ICS.

R	Aggregation	TP	FP	CR	MSE
50	lasso	4.00 (0.00)	0.51 (1.08)	1.00 (0.00)	0.031 (0.019)
	SCAD	4.00 (0.00)	0.00 (0.00)	1.00 (0.00)	0.028 (0.017)
	Averaging	4.00 (0.00)	100.97 (22.55)	1.00 (0.00)	0.028 (0.017)
100	lasso	4.00 (0.00)	0.38 (0.56)	1.00 (0.00)	0.031 (0.019)
	SCAD	4.00 (0.00)	0.00 (0.00)	1.00 (0.00)	0.027 (0.017)
	Averaging	4.00 (0.00)	160.00 (30.16)	1.00 (0.00)	0.028 (0.017)
200	lasso	4.00 (0.00)	0.34 (0.62)	1.00 (0.00)	0.031 (0.018)
	SCAD	4.00 (0.00)	0.00 (0.00)	1.00 (0.00)	0.027 (0.016)
	Averaging	4.00 (0.00)	231.82 (37.95)	1.00 (0.00)	0.027 (0.016)
500	lasso	4.00 (0.00)	0.19 (0.44)	1.00 (0.00)	0.032 (0.018)
	SCAD	4.00 (0.00)	0.00 (0.00)	1.00 (0.00)	0.028 (0.016)
	Averaging	4.00 (0.00)	338.77 (41.11)	1.00 (0.00)	0.028 (0.016)
1000	lasso	4.00 (0.00)	0.14 (0.38)	1.00 (0.00)	0.033 (0.019)
	SCAD	4.00 (0.00)	0.00 (0.00)	1.00 (0.00)	0.028 (0.017)
	Averaging	4.00 (0.00)	410.62 (35.41)	1.00 (0.00)	0.028 (0.017)

Table 3. Model selection results for Example 2 with ICS.

p_n	ρ	Approach	TP	FP	CR	MSE
50	0.5	PL_{WCR}	3.00 (0.00)	0.26 (0.50)	1.00 (0.00)	0.040 (0.052)
		PGEE.indep	1.77 (1.48)	11.81 (10.18)	0.59 (0.49)	1.043 (1.055)
		PGEE.exch	3.00 (0.00)	23.57 (3.72)	1.00 (0.00)	0.136 (0.060)
		PGEE.ar1	2.85 (0.66)	20.43 (5.54)	0.95 (0.22)	0.256 (0.476)
		naive lasso	1.94 (0.78)	0.00 (0.00)	0.27 (0.45)	1.698 (0.152)
	0.8	PL_{WCR}	3.00 (0.00)	0.23 (0.45)	1.00 (0.00)	0.048 (0.051)
		PGEE.indep	1.65 (1.50)	11.10 (10.44)	0.55 (0.50)	1.142 (1.056)
		PGEE.exch	3.00 (0.00)	25.46 (3.91)	1.00 (0.00)	0.114 (0.049)
		PGEE.ar1	2.97 (0.30)	23.43 (4.46)	0.99 (0.10)	0.156 (0.225)
		naive lasso	1.83 (0.75)	0.00 (0.00)	0.21 (0.41)	1.734 (0.152)
500	0.5	PL_{WCR}	3.00 (0.00)	0.03 (0.17)	1.00 (0.00)	0.101 (0.075)
		naive lasso	1.90 (0.69)	0.01 (0.10)	0.19 (0.39)	1.713 (0.132)
	0.8	PL_{WCR}	3.00 (0.00)	0.03 (0.17)	1.00 (0.00)	0.115 (0.096)
		naive lasso	1.81 (0.69)	0.00 (0.00)	0.16 (0.37)	1.729 (0.138)

In Examples 3 and 4, we intentionally exclude ICS in order to evaluate the robustness of the proposed PL_{WCR} method under scenarios where ICS is not a factor. This design allows us to isolate and examine the method's performance in handling high-dimensional longitudinal data without the confounding influence of ICS, providing a clearer assessment of its general applicability and stability. By focusing on these examples, we aim to demonstrate that our PL_{WCR} method remains effective even in the absence of ICS, thereby further validating its utility across diverse longitudinal data settings.

Example 3. *We consider the linear model*

$$Y_{ij} = \boldsymbol{X}_{ij}^T \boldsymbol{\beta} + \epsilon_{ij}$$

for correlated normal responses, where $i = 1, \ldots, 200$, $j = 1, \ldots, M_i$. *The values of* M_i *are generated following the same procedure as in Example 1. The coefficient vector is specified as* $\boldsymbol{\beta} = (2, -1, 1.5, -2, 0, 0, \ldots, 0)^T$, *representing a homogeneous effect. The covariates are generated according to the method described in Example 1. The random errors* $(\epsilon_{i1}, \ldots, \epsilon_{iM_i})^T$ *follow a multivariate normal distribution with marginal mean 0, marginal variance 1, and exchangeable correlation matrix characterized by a parameter* $\rho = 0.5$ or 0.8.

Table 4. Sensitivity analyses for PL_{WCR} with alternative aggregation for Example 2 with ICS.

R	Aggregation	TP	FP	CR	MSE
50	lasso	3.00 (0.00)	0.25 (0.52)	1.00 (0.00)	0.117 (0.008)
	SCAD	3.00 (0.00)	0.00 (0.00)	1.00 (0.00)	0.124 (0.107)
	Averaging	3.00 (0.00)	477.32 (5.69)	1.00 (0.00)	0.273 (0.109)
100	lasso	3.00 (0.00)	0.18 (0.50)	1.00 (0.00)	0.105 (0.074)
	SCAD	3.00 (0.00)	0.00 (0.00)	1.00 (0.00)	0.110 (0.099)
	Averaging	3.00 (0.00)	496.08 (0.99)	1.00 (0.00)	0.232 (0.093)
200	lasso	3.00 (0.00)	0.11 (0.37)	1.00 (0.00)	0.107 (0.077)
	SCAD	3.00 (0.00)	0.00 (0.00)	1.00 (0.00)	0.109 (0.099)
	Averaging	3.00 (0.00)	497.00 (0.00)	1.00 (0.00)	0.219 (0.094)
500	lasso	3.00 (0.00)	0.03 (0.17)	1.00 (0.00)	0.101 (0.075)
	SCAD	3.00 (0.00)	0.00 (0.00)	1.00 (0.00)	0.101 (0.094)
	Averaging	3.00 (0.00)	497.00 (0.00)	1.00 (0.00)	0.204 (0.090)
1000	lasso	3.00 (0.00)	0.01 (0.10)	1.00 (0.00)	0.100 (0.074)
	SCAD	3.00 (0.00)	0.00 (0.00)	1.00 (0.00)	0.100 (0.092)
	Averaging	3.00 (0.00)	497.00 (0.00)	1.00 (0.00)	0.203 (0.089)

Table 5 summarizes the model selection and estimation results for continuous correlated responses in Example 3. Overall, the proposed PL_{WCR} method successfully identifies all relevant covariates while achieving the most parsimonious model size among the competing methods. Under the data generation mechanism specified in Example 3, the PGEE method also provides a sparse selected model and optimal estimation, particularly when assuming a pairwise equal correlation structure. This is because the generalized estimating equations (GEE) approach provides consistent and efficient estimates of regression coefficients while accounting for within-cluster dependencies. GEE utilizes all available observations and adjusts for intracluster correlation. When the a priori specified working correlation structure approximates the true dependence, the GEE estimator achieves optimal efficiency, minimizing the asymptotic variance of parameter estimates. Our simulation results demonstrate that GEE maintains type I error control and statistical power. In comparison, the MSEs of the PL_{WCR} method are comparable to those of PGEE, indicating that our method also enjoys robust performance regardless of ICS. In contrast, naive LASSO selects moderately sized models but produces biased estimates, with the bias becoming more pronounced at $p_n = 500$.

When combined with the results from Example 1, it is apparent that the proposed PL_{WCR} method demonstrates robust model selection performance and accurate estimation regardless of whether ICS is ignorable. These findings highlight the versatility and reliability of our PL_{WCR} approach in handling high-dimensional longitudinal data across diverse scenarios.

In Table 6, it can be seen that the results of the sensitivity analyses for Example 3 are similar to those in Example 1 at $\rho = 0.5$ and $p_n = 500$. After eliminating the influence of

ICS, all three aggregation methods exhibit significant improvements in estimation efficiency. Although simple averaging substantially reduces the FPs in the absence of nonignorable cluster sizes, the selected model remained non-sparse. The model selection performance of the linear models exhibits little sensitivity to variations in the sampling frequency.

Table 5. Model selection results for Example 3 without ICS.

p_n	ρ	Approach	TP	FP	CR	MSE
50	0.5	PL_{WCR}	4.00 (0.00)	0.60 (0.79)	1.00 (0.00)	0.010 (0.010)
		PGEE.indep	4.00 (0.00)	2.37 (3.56)	1.00 (0.00)	0.011 (0.010)
		PGEE.exch	4.00 (0.00)	4.90 (5.50)	1.00 (0.00)	0.008 (0.008)
		PGEE.ar1	4.00 (0.00)	4.65 (4.79)	1.00 (0.00)	0.008 (0.006)
		naive lasso	4.00 (0.00)	2.49 (2.11)	1.00 (0.00)	0.097 (0.042)
	0.8	PL_{WCR}	4.00 (0.00)	0.51 (0.72)	1.00 (0.00)	0.010 (0.009)
		PGEE.indep	4.00 (0.00)	2.33 (3.21)	1.00 (0.00)	0.009 (0.008)
		PGEE.exch	4.00 (0.00)	9.91 (12.92)	1.00 (0.00)	0.009 (0.017)
		PGEE.ar1	4.00 (0.00)	12.69 (11.75)	1.00 (0.00)	0.008 (0.009)
		naive lasso	4.00 (0.00)	2.18 (1.58)	1.00 (0.00)	0.095 (0.036)
500	0.5	PL_{WCR}	4.00 (0.00)	0.19 (0.46)	1.00 (0.00)	0.009 (0.006)
		naive lasso	4.00 (0.00)	3.09 (2.04)	1.00 (0.00)	0.207 (0.062)
	0.8	PL_{WCR}	4.00 (0.00)	0.25 (0.50)	1.00 (0.00)	0.009 (0.006)
		naive lasso	4.00 (0.00)	3.53 (2.43)	1.00 (0.00)	0.204 (0.064)

Example 4. *In this example, we model the correlated binary responses with the marginal mean defined as*

$$E(Y_{ij} \mid \boldsymbol{X}_{ij}, M_i) = \exp(\boldsymbol{X}_{ij}^T\boldsymbol{\beta})/\{1 + \exp(\boldsymbol{X}_{ij}^T\boldsymbol{\beta})\},$$

where $i = 1, \ldots, 200$ *and* $j = 1, \ldots, M_i$. *The random cluster sizes and covariates are generated following the procedure described in Example 2. The coefficient vector is set as* $\boldsymbol{\beta} = (1, -0.9, 0.7, 0, 0, \ldots, 0)^T$. *The observations within each cluster exhibit an exchangeable correlation structure with parameter* $\rho = 0.5$ or 0.8.

Table 7 provides a summary of the model selection results and estimation accuracy for the competing methods in Example 4. When the correlated binary responses are independent of the cluster size, our proposed PL_{WCR} method consistently outperforms alternative approaches in both signal identification and parameter estimation. While the PGEE method exhibits instability in terms of TPs and FPs across different working correlation matrices, it demonstrates improvement in identifying significant features compared to its performance in Example 2. Although the naive LASSO method is able to achieve model sparsity, its performance is inferior to the proposed approach due to its biased parameter estimation. As the number of covariates increases to $p_n = 500$, our PL_{WCR} method continues to excel, selecting an oracle model with the smallest MSE. These results underscore the robustness and precision of our PL_{WCR} method in handling high-dimensional longitudinal data with binary responses even in the absence of ICS.

Table 8 summarizes the results of the sensitivity analyses for Example 4 at $\rho = 0.5$ and $p_n = 500$. While increasing R reduces parameter estimation errors, this improvement comes at the cost of elevated computational demands. Notably, our experiments demonstrate nearly equivalent performance between $R = 500$ and $R = 1000$, suggesting diminishing returns beyond $R = 500$. When combined, the results from Tables 2–8 show that 500 times resampling for our PL_{WCR} method achieves good robustness and tradeoffs between computational efficiency and stability.

Table 6. Sensitivity analyses for PL_{WCR} with alternative aggregation for Example 3 without ICS.

R	Aggregation	TP	FP	CR	MSE
50	lasso	4.00 (0.00)	0.23 (0.49)	1.00 (0.00)	0.010 (0.007)
	SCAD	4.00 (0.00)	0.00 (0.00)	1.00 (0.00)	0.010 (0.007)
	Averaging	4.00 (0.00)	39.35 (11.79)	1.00 (0.00)	0.010 (0.007)
100	lasso	4.00 (0.00)	0.19 (0.42)	1.00 (0.00)	0.010 (0.007)
	SCAD	4.00 (0.00)	0.00 (0.00)	1.00 (0.00)	0.009 (0.007)
	Averaging	4.00 (0.00)	64.24 (17.21)	1.00 (0.00)	0.010 (0.007)
200	lasso	4.00 (0.00)	0.19 (0.44)	1.00 (0.00)	0.009 (0.007)
	SCAD	4.00 (0.00)	0.00 (0.00)	1.00 (0.00)	0.009 (0.007)
	Averaging	4.00 (0.00)	99.87 (23.70)	1.00 (0.00)	0.009 (0.006)
500	lasso	4.00 (0.00)	0.19 (0.46)	1.00 (0.00)	0.009 (0.006)
	SCAD	4.00 (0.00)	0.00 (0.00)	1.00 (0.00)	0.009 (0.006)
	Averaging	4.00 (0.00)	164.24 (33.09)	1.00 (0.00)	0.009 (0.006)
1000	lasso	4.00 (0.00)	0.15 (0.41)	1.00 (0.00)	0.009 (0.007)
	SCAD	4.00 (0.00)	0.00 (0.00)	1.00 (0.00)	0.009 (0.007)
	Averaging	4.00 (0.00)	223.06 (40.78)	1.00 (0.00)	0.009 (0.007)

Table 7. Model selection results for Example 4 without ICS.

p_n	ρ	Approach	TP	FP	CR	MSE
50	0.5	PL_{WCR}	3.00 (0.00)	0.34 (0.57)	1.00 (0.00)	0.049 (0.057)
		PGEE.indep	2.10 (1.38)	14.09 (9.70)	0.70 (0.46)	0.813 (0.980)
		PGEE.exch	2.97 (0.30)	20.57 (3.66)	0.99 (0.10)	0.173 (0.225)
		PGEE.ar1	2.88 (0.59)	20.19 (5.28)	0.96 (0.20)	0.249 (0.426)
		naive lasso	2.29 (0.82)	0.00 (0.00)	0.52 (0.50)	1.547 (0.214)
	0.8	PL_{WCR}	3.00 (0.00)	0.25 (0.46)	1.00 (0.00)	0.043 (0.052)
		PGEE.indep	1.95 (1.44)	1.09 (3.70)	0.65 (0.48)	0.926 (1.015)
		PGEE.exch	3.00 (0.00)	2.94 (1.52)	1.00 (0.00)	0.111 (0.048)
		PGEE.ar1	3.00 (0.00)	2.87 (2.75)	1.00 (0.00)	0.138 (0.062)
		naive lasso	2.25 (0.80)	0.01 (0.10)	0.47 (0.50)	1.582 (0.198)
500	0.5	PL_{WCR}	3.00 (0.00)	0.02 (0.14)	1.00 (0.00)	0.086 (0.066)
		naive lasso	2.27 (0.76)	0.00 (0.00)	0.46 (0.50)	1.587 (0.203)
	0.8	PL_{WCR}	3.00 (0.00)	0.06 (0.28)	1.00 (0.00)	0.091 (0.065)
		naive lasso	2.20 (0.78)	0.01 (0.10)	0.42 (0.50)	1.615 (0.212)

Table 8. Sensitivity analyses for PL_{WCR} with alternative aggregation for Example 4 without ICS.

R	Aggregation	TP	FP	CR	MSE
50	lasso	3.00 (0.00)	0.23 (0.49)	1.00 (0.00)	0.104 (0.083)
	SCAD	3.00 (0.00)	0.00 (0.00)	1.00 (0.00)	0.120 (0.097)
	Averaging	3.00 (0.00)	478.82 (6.18)	1.00 (0.00)	0.279 (0.091)
100	lasso	3.00 (0.00)	0.20 (0.53)	1.00 (0.00)	0.097 (0.078)
	SCAD	3.00 (0.00)	0.00 (0.00)	1.00 (0.00)	0.106 (0.090)
	Averaging	3.00 (0.00)	495.91 (1.16)	1.00 (0.00)	0.231 (0.079)
200	lasso	3.00 (0.00)	0.19 (1.03)	1.00 (0.00)	0.092 (0.074)
	SCAD	3.00 (0.00)	0.00 (0.00)	1.00 (0.00)	0.097 (0.081)
	Averaging	3.00 (0.00)	497.00 (0.00)	1.00 (0.00)	0.213 (0.081)
500	lasso	3.00 (0.00)	0.02 (0.14)	1.00 (0.00)	0.086 (0.066)
	SCAD	3.00 (0.00)	0.00 (0.00)	1.00 (0.00)	0.088 (0.065)
	Averaging	3.00 (0.00)	497.00 (0.00)	1.00 (0.00)	0.194 (0.058)
1000	lasso	3.00 (0.00)	0.01 (0.10)	1.00 (0.00)	0.084 (0.064)
	SCAD	3.00 (0.00)	0.00 (0.00)	1.00 (0.00)	0.086 (0.065)
	Averaging	3.00 (0.00)	497.00 (0.00)	1.00 (0.00)	0.189 (0.060)

5. Yeast Cell-Cycle Gene Expression Data Analysis

In this section, we assess the performance of our proposed PL_{WCR} method using a subset of the yeast cell-cycle gene expression dataset compiled by [22]. This dataset comprises 292 cell-cycle-regulated genes with expression levels monitored across two complete cell-cycle periods. Repeated measurements were collected for these 292 genes at intervals of 7 min over a span of 119 min, yielding a total of 18 time points. This dataset provides a valuable opportunity to evaluate the efficacy of our method in analyzing high-dimensional longitudinal data with inherent temporal dependencies.

The cell-cycle process is typically divided into a number of distinct stages: M/G1, G1, S, G2, and M. The M (mitosis) stage involves nuclear events such as chromosome separation as well as cytoplasmic events such as cytokinesis and cell division. The G1 (GAP 1) stage precedes DNA synthesis, while the S (synthesis) stage is characterized by DNA replication. The G2 (GAP 2) stage follows synthesis, and prepares the cell for mitosis. Transcription factors (TFs) play a crucial role in regulating the transcription of a subset of yeast cell-cycle-regulated genes. In [23], the authors utilized ChIP data from [24] to estimate the binding probabilities for these 292 genes, covering a total of 96 TFs.

To identify the TFs that are potentially involved in the yeast cell-cycle, we consider the following model:

$$y_{ij} = \sum_{d=1}^{96} \beta^*_{j,d} x_{ij,d} + \sum_{l_1,l_2=1,\ldots,96, l_1<l_2} \gamma^*_{j,l_1l_2} x_{ij,l_1} x_{ij,l_2} + \epsilon_{ij} \tag{9}$$

where y_{ij} represents the log-transformed gene expression level of gene i measured at time point j for $i = 1, \ldots, 292$ and $j = 1, \ldots, 18$. The variable $x_{ij,d}$ denotes the matching score of the binding probability of the dth TF for gene i at time point j. Previous studies have highlighted the importance of interaction effects among certain TFs, such as HIR1:HIR2 [25], SWI4:SWI6 [26], FKH2:NDD1 [27], SWI5:SMP1 [28], MCM1:FKH2 [29], MBP1:SWI6 [26], and FKH1:FKH2 [29,30]. Given this evidence, the model in (9) is well justified, resulting in a total of 4656 covariates for analysis.

As discussed in Section 4, we compare the variable selection performance of the proposed PL_{WCR} method with that of the PGEE and naive LASSO methods. Table 9 summarizes the identification results for 21 transcription factors (TFs) that have been experimentally verified and are known to be associated with the yeast cell cycle (referred to as "true" TFs) [23]. This comparison highlights the ability of each method to accurately identify biologically relevant TFs, providing insights into their respective strengths and limitations in the context of high-dimensional longitudinal data analysis. The proposed PL_{WCR} method demonstrates superior performance in identifying true TFs while maintaining model sparsity, further validating its utility in biological applications. Overall, the proposed PL_{WCR} method selects the smallest number of TFs while identifying 17 "true" TFs, which is the highest count among all competing methods. In contrast, PGEE exhibits variable performance depending on the assumed within-cluster correlation structures. When observations are modeled as independent or equally correlated, PGEE identifies 11 "true" TFs out of 60 and 61 selected TFs, respectively. Under an autoregressive correlation structure, PGEE improves slightly, identifying 15 "true" TFs among 60 selected TFs. These results underscore the strong dependence of PGEE's performance on the choice of working correlation structure. Meanwhile, the naive LASSO method identifies 12 "true" TFs from a comparable total number of selected TFs. These findings highlight the robustness and precision of our PL_{WCR} method in accurately identifying biologically relevant TFs while maintaining model sparsity, even in the presence of complex correlation structures.

Next, we examine the identification of specific "true" TFs in more detail. All of the competing methods successfully select ACE2, CBF1, FKH1, MBP1, NDD1, REB1, STB1, SWI4, and SWI5.

- BAS1: BAS1 is known to regulate the synthesis of histidine, purines, and pyrimidines [31], and also plays a role in preparing yeast cells for division. However, the penalized GEE method fails to identify it as important.
- MCM1: This TF primarily exerts its regulatory effects during the M/G1, S, and early G2 stages, indicating its critical role in the yeast cell cycle. Although it is identified as one of the three major TF types by [22], MCM1's vital impact is overlooked by the penalized GEE method.
- MET31: MET31 is required for interactions with the activator MET4 to bind DNA. It regulates sulfur metabolism [32], a process linked to the initiation of cell division [33]. Notably, only our PL_{WCR} method identifies the importance of MET31.

Despite achieving sparsity, the naive LASSO method fails to select several significant TFs with substantial biological roles:

- FKH2, which plays a dominant role in regulating genes associated with nuclear migration and spindle formation [34].
- SWI6, which coordinates gene expression at the G1-S boundary of the yeast cell cycle, as noted by [35].

These results highlight the superior performance of the proposed PL_{WCR} method in reliably identifying critical TFs associated with the yeast cell cycle. However, a few TFs were not successfully detected by PL_{WCR}. Figures 1–4 show scatter plots of the matching score of the binding probability of four missing "true" TFs: GCN4, GCR1, SKN7, and STE12, respectively. A common characteristic can be observed among these TFs in that the majority of observations are clustered near zero, with only a limited number of outliers exhibiting deviations. This phenomenon results in our resampling-based penalized estimation approach misclassifying these TFs as invariant constants rather than as variables, causing their exclusion from the identified set. In particular, GCR1 displayed an exceptionally low variance (on the order of 10^{-6}), effectively approximating a constant zero expression level.

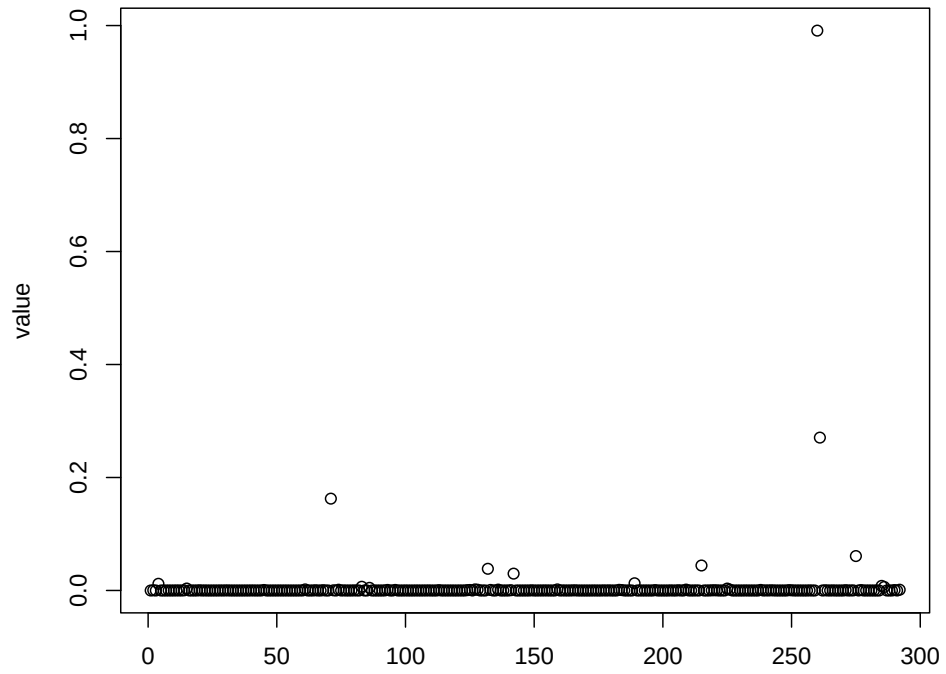


Figure 1. Matching score of GCN4.

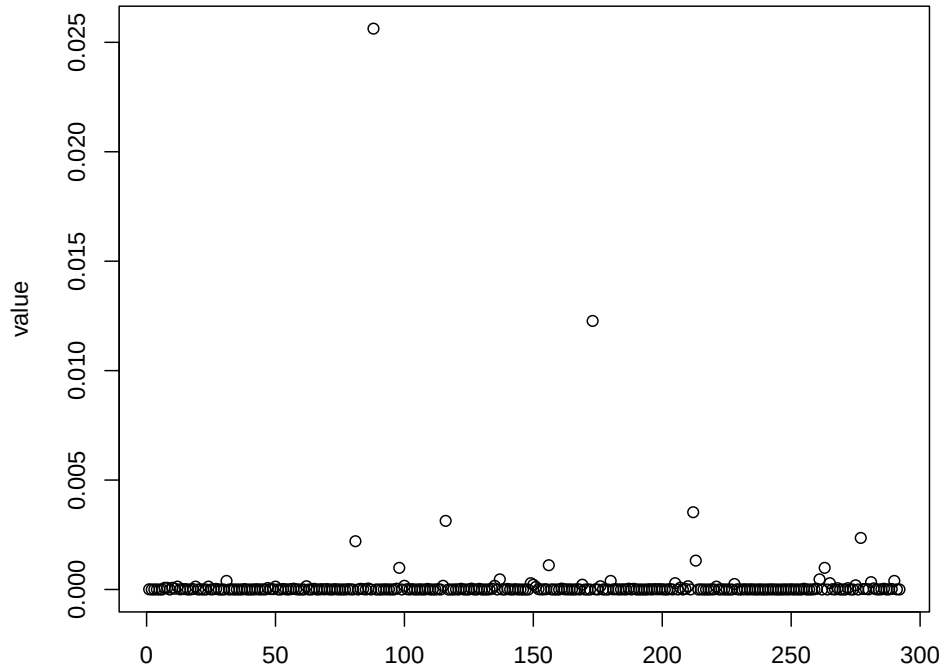


Figure 2. Matching score of GCR1.

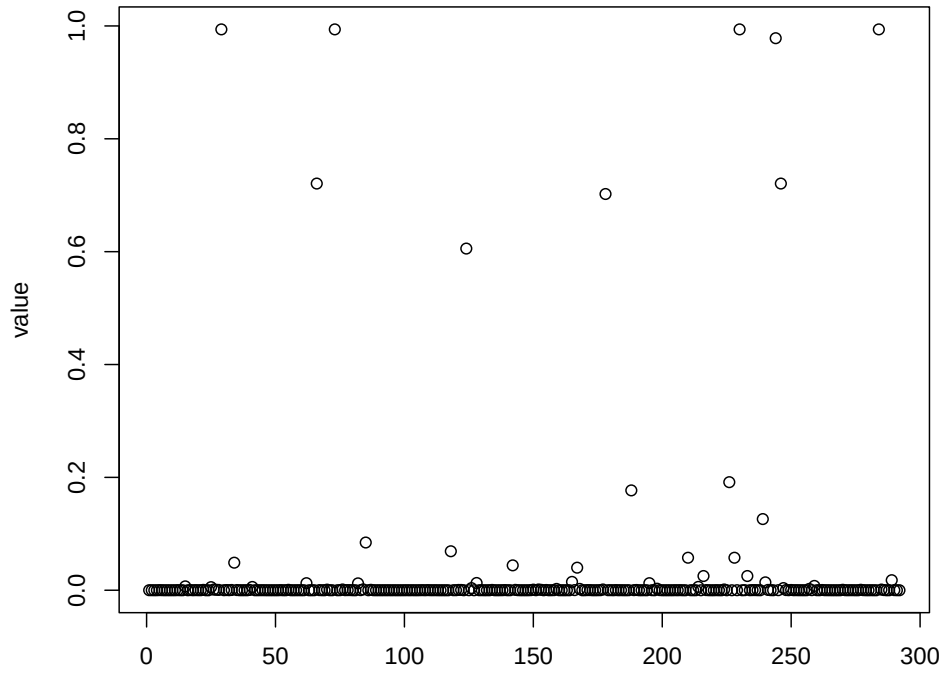


Figure 3. Matching score of SKN7.

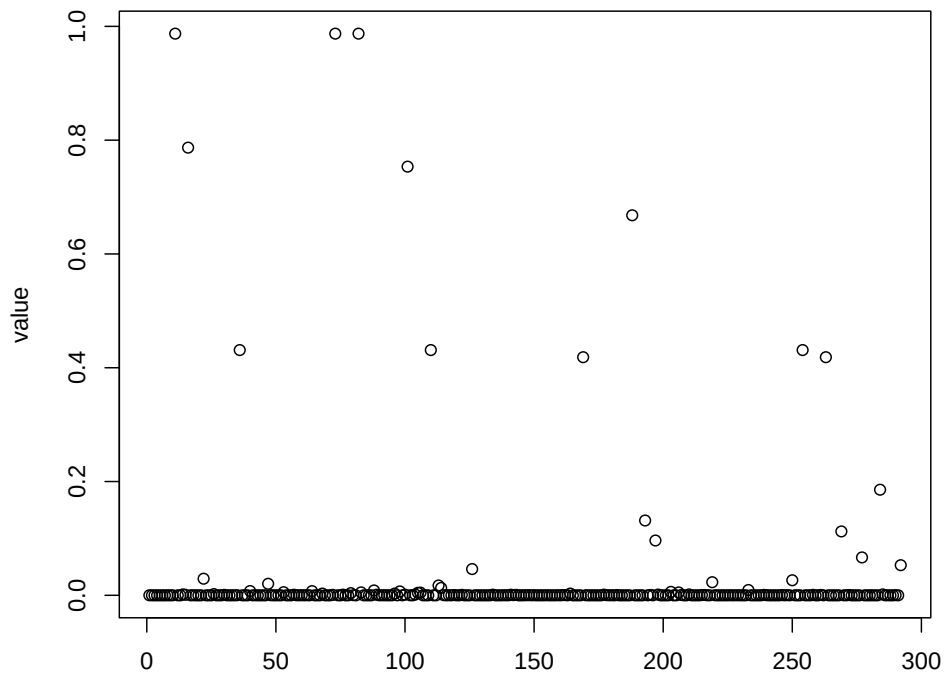


Figure 4. Matching score of STE12.

Notably, GCN4 was not identified as significant by any of the considered methods. The naive LASSO method was the only approach to select GCR1, while SKN7 and STE12 were recognized as cell cycle regulators exclusively by the PGEE method. To better understand these discrepancies, we turn to relevant biological research for plausible explanations. First, GCN4, which was not detected by any method, may play a regulatory role under specific conditions or stress responses that are not fully captured in the current dataset. Similarly, GCR1 was identified only by naive LASSO, and might exhibit weak or context-dependent regulatory effects that are more challenging to detect using methods that account for within-cluster correlations. The identification of SKN7 and STE12 by PGEE but not by PL_{WCR} could be attributed to the sensitivity of PGEE to specific correlation structures, which may align more closely with the regulatory patterns of these TFs.

Relevant biological research can offer plausible explanations that may aid in better understanding these discrepancies. First, the heterogeneous regulatory effects of TFs during the yeast cell cycle likely contribute to these observations. Different TFs exhibit varying levels of regulatory involvement at distinct stages of the cycle. Furthermore, certain TFs regulate the cell cycle in a periodic manner, with their effects confined to specific intervals. Because our PL_{WCR} method samples observations at arbitrary time points, it tends to prioritize TFs with consistent regulatory effects across the two observed periods. For example, STE12 is known to regulate the cyclic expression of certain genes specifically during the early G1 phase, as documented by [36]. This phase represents a relatively narrow time window, resulting in a low sampling probability and an estimated coefficient close to zero. However, if observations were specifically sampled during the G1 phase, STE12 could be successfully identified. This highlights the importance of considering temporal dynamics and stage-specific regulatory mechanisms when analyzing cell-cycle-related gene expression data.

Table 9. Results for identifying the 21 "true" TFs; "1" means that the TF is selected, while "0" means the opposite.

TF	PL_{WCR}	PGEE.indep	PGEE.exch	PGEE.ar1	Naive Lasso
ABF1	1	1	1	1	0
ACE2	1	0	0	1	1
BAS1	1	0	0	0	1
CBF1	1	1	1	1	1
FKH1	1	1	1	1	1
FKH2	1	1	1	1	0
GCN4	0	0	0	0	0
GCR1	0	0	0	0	1
GCR2	1	0	0	1	0
LEU3	1	1	1	0	0
MBP1	1	0	0	1	1
MCM1	1	0	0	0	1
MET31	1	0	0	0	0
NDD1	1	1	1	1	1
REB1	1	1	1	1	1
SKN7	0	0	0	1	0
STB1	1	1	1	1	1
STE12	0	1	1	1	0
SWI4	1	0	0	1	1
SWI5	1	1	1	1	1
SWI6	1	1	1	1	0
Number of "true" TFs	17	11	11	15	12
Total number of TFs	52	60	61	60	60

Additionally, the complex interactions between transcription factors (TFs) may influence the identification results. For instance, SKN7 exhibits several genetic interactions with MBP1 during the G1-S transition, including mutual inhibition [37]. Despite this interaction, our PL_{WCR} method identified MBP1 as significant, potentially overshadowing SKN7. This suggests that the regulatory influence of MBP1 as captured by the model may dominate or mask the effects of SKN7 due to their antagonistic relationship. Such intricate interaction between TFs underscores the challenges in disentangling their individual contributions and highlights the need for more sophisticated modeling approaches.

In addition to the 21 experimentally verified TFs, our PL_{WCR} method also identified additional regulatory TFs with biological evidence supporting their roles in the yeast cell cycle. For example, HIR1 is known to be involved in the cell cycle-dependent repression of histone gene transcription [38]. Similarly, [39] demonstrated that PHD1 contributes to centriole duplication and centrosome maturation, playing a critical role in regulating cell cycle progression. These findings not only validate the accuracy of our PL_{WCR} method in identifying biologically relevant TFs but also highlight its potential to uncover additional regulatory factors that may have been overlooked by other approaches. This underscores our proposed method's utility in advancing the understanding of complex biological processes such as cell cycle regulation.

6. Discussion

In this article, we have addressed the challenges of variable selection and parameter estimation in high-dimensional longitudinal data with ICS. By leveraging the WCR technique, we develop a penalized likelihood (PL) method applied to each resampled dataset. Theoretical analysis establishes the consistency of both model selection and parameter estimation for this approach. To integrate the estimators derived from multiple resampled

datasets, we introduce a penalized mean regression technique, which produces an aggregated estimator that enhances the identification of true positives while minimizing false positives. This integrated methodology, which we term PL_{WCR}, combines the strengths of WCR and PL. Simulation studies demonstrate that PL_{WCR} achieves consistent model selection and robust estimation even when outcomes are dependent on cluster size.

The performance of PL_{WCR} is fundamentally linked to the underlying resampling mechanism. While WCR effectively mitigates bias arising from informative cluster size (ICS), the stability and variability of the resulting estimates are heavily influenced by the resampling design. Using smaller resample sizes can increase the Monte Carlo error, whereas excessively large resample sizes may reduce computational efficiency without offering commensurate gains in precision. Another critical aspect of PL_{WCR} is its dependence on penalization parameters, which control both sparsity and the aggregation of estimators. Although theoretical guarantees such as consistency hold in asymptotic settings, finite-sample performance remains sensitive to the selection of tuning parameters. While cross-validation is widely used in practice, it may become unstable in high-dimensional settings or when the signal-to-noise ratio is low. Furthermore, the penalized mean regression step assumes that the weighted contributions from resampled estimates are optimal, an assumption that may not always hold in practice.

Despite these limitations, PL_{WCR} provides an efficient and reliable analytical tool for longitudinal data, requiring only first- and second-moment assumptions. This approach represents a significant advancement in handling high-dimensional longitudinal data with ICS, offering both theoretical guarantees and practical utility. Future work could explore adaptive resampling strategies or data-driven penalization to further enhance robustness.

Author Contributions: Conceptualization, X.J.; Methodology, Y.M. and X.J.; Software, Y.M.; Formal analysis, Y.M.; Investigation, Y.M.; Writing—original draft, Y.M.; Supervision, X.J.; Project administration, X.J. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by National Natural Science Foundation of China (12271238) and Guang Dong Basic and Applied Basic Research Foundation (2023A1515010025) for Xuejun Jiang.

Data Availability Statement: The yeast cell-cycle gene expression dataset used in this article are from [22], further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Proof of Theorem 1. For simplicity, we let $\mathbf{X_z} = (\mathbf{X}_{1z_1}^T, \ldots \mathbf{X}_{nz_n}^T)^T$ and let $\mathbf{X_{z1}}$ and $\mathbf{X_{z2}}$ be the submatrices of $\mathbf{X_z}$ formed by columns in $\mathcal{S}$ and $\mathcal{S}^c$, where $\mathcal{S}$ is the index set corresponding to nonzero components in $\boldsymbol{\beta}^*$. Let $\boldsymbol{\beta}_1$ be the corresponding coefficient vector to $\mathbf{X}_1$. Define $\boldsymbol{\varepsilon}_\mathbf{z}(\boldsymbol{\beta}) = (\mathbf{Y_z} - \mathbf{b}'(\mathbf{X_z}\boldsymbol{\beta}))$ and $\bar{\boldsymbol{\rho}}_{\lambda_n}(\mathbf{v}) = (sgn(v_1)\rho'_{\lambda_n}(|v_1|), \ldots, sgn(v_q)\rho'_{\lambda_n}(|v_q|))^T$, where $\mathbf{v} = (v_1, \ldots, v_q)^T$, $\mathbf{Y_z} = (Y_{1z_1}, \ldots, Y_{nz_n})^T$. Denote $\mathbf{I}_n(\boldsymbol{\beta}_1; \mathbf{z}) = -\partial^2 \tilde{L}_n(\boldsymbol{\beta}_1, \mathbf{0}; \mathbf{z})/\partial\boldsymbol{\beta}_1\partial\boldsymbol{\beta}_1^T$. Similar to the proof of Theorem 3 in [17], it is sufficient to show that the following hold with probability going to one:

(i) (Consistency) There exists a strict local maximizer $\hat{\boldsymbol{\beta}}_1$ of $Q_n(\mathbf{a}, \mathbf{0}; \mathbf{z})$ such that $\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_2 = O_P(\sqrt{s_n/n})$;
(ii) (Sparsity) $\|(n\lambda_n)^{-1}\mathbf{X}_{\mathbf{z}2}^T\boldsymbol{\varepsilon}_\mathbf{z}(\hat{\boldsymbol{\beta}}_1, \mathbf{0})\}\|_\infty < \rho'(0+)$;
(iii) $\lambda_{\min}(\mathbf{I}_n(\hat{\boldsymbol{\beta}}_1; \mathbf{z})) > \lambda_n\kappa(\rho; \hat{\boldsymbol{\beta}}_1)$.

(i) Define the boundary set as

$$\mathcal{M} = \{\mathbf{a} : \|\mathbf{a} - \boldsymbol{\beta}_1^*\|_2 = \tau\sqrt{s_n/n}\}$$

for $\tau > 0$. Application of Taylor expansion to $\ell_n(\mathbf{a}, \mathbf{0}; \mathbf{z})$ at $\boldsymbol{\beta}^*$ yields

$$\begin{aligned} Q_n(\mathbf{a}, \mathbf{0}; \mathbf{z}) - Q_n(\boldsymbol{\beta}^*) =& (\mathbf{a} - \boldsymbol{\beta}_1^*)^T \{n^{-1}\mathbf{X}_{\mathbf{z}1}^T \boldsymbol{\varepsilon}_{\mathbf{z}}(\boldsymbol{\beta}^*) - \lambda_n \bar{\boldsymbol{\rho}}_{\lambda_n}(\boldsymbol{\beta}_1^*)\} \\ & - \frac{1}{2}(\mathbf{a} - \boldsymbol{\beta}_1^*)^T \{\mathbf{I}_{1n}(\tilde{\boldsymbol{\beta}}_1; \mathbf{z}) + \mathrm{diag}(\lambda_n \rho''(|\tilde{\boldsymbol{\beta}}_1|))\}(\mathbf{a} - \boldsymbol{\beta}_1^*), \end{aligned}$$

where $\tilde{\boldsymbol{\beta}}_1$ lies on the line segment joining $\mathbf{a}$ and $\boldsymbol{\beta}_1^*$. For any $\mathbf{a} \in \mathcal{M}$, by $\sqrt{s_n/n} = o(d_n)$ in Condition (A4), we obtain

$$\min_{1 \le j \le s_n} |\tilde{\boldsymbol{\beta}}_{1,j}| \ge \min_{j \in \mathcal{S}} |\beta_j^*| - \tau\sqrt{s_n/n} \ge d_n$$

for sufficiently large n. Then, by the monotonicity condition in Condition (A1), it follows that

$$\lambda_n \|\bar{\rho}(\boldsymbol{\beta}_1^*)\|_\infty \le \lambda_n \rho'_{\lambda_n}(d_n), \tag{A1}$$

which, combined with Condition (A2), leads to

$$\begin{aligned} & Q_n(\mathbf{a}, \mathbf{0}; \mathbf{z}) - Q_n(\boldsymbol{\beta}^*) \\ \le & n^{-1}\|\mathbf{a} - \boldsymbol{\beta}_1^*\|_2 \|\mathbf{X}_{\mathbf{z}1}^T \boldsymbol{\varepsilon}_{\mathbf{z}}(\boldsymbol{\beta}^*)\|_2 - \frac{1}{2}(\lambda_{\min}(\mathbf{I}_{1n}(\tilde{\boldsymbol{\beta}}_1; \mathbf{z})) + \lambda_n \kappa(\rho_{\lambda_n}))\|\mathbf{a} - \boldsymbol{\beta}_1^*\|_2^2 + \tau s_n n^{-1/2} \lambda_n \rho'_{\lambda_n}(d_n) \\ = & n^{-1}\tau\sqrt{s_n/n}\|\mathbf{X}_{\mathbf{z}1}^T \boldsymbol{\varepsilon}_{\mathbf{z}}(\boldsymbol{\beta}^*)\|_2 - 0.5\tau^2 s_n n^{-1}(\lambda_{\min}(\mathbf{I}_{1n}(\tilde{\boldsymbol{\beta}}_1; \mathbf{z})) + o(1)) + \tau s_n n^{-1/2} \lambda_n \rho'_{\lambda_n}(d_n) \\ \le & n^{-1}\tau\sqrt{s_n/n}\|\mathbf{X}_{\mathbf{z}1}^T \boldsymbol{\varepsilon}_{\mathbf{z}}(\boldsymbol{\beta}^*)\|_2 - 0.5c_3\tau^2 s_n n^{-1} + \tau s_n n^{-1/2} \lambda_n \rho'_{\lambda_n}(d_n) \end{aligned}$$

holding uniformly for any $\mathbf{a} \in \mathcal{M}$. Here, the last inequality follows from Condition (A3). By Condition (A2), we have

$$\begin{aligned} E\|\mathbf{X}_{\mathbf{z}1}^T \boldsymbol{\varepsilon}_{\mathbf{z}}\|_2^2 &= \mathrm{Trace}\{E(\mathbf{X}_{\mathbf{z}1}^T \mathrm{diag}(\boldsymbol{\varepsilon}_{\mathbf{z}}^2)\mathbf{X}_{\mathbf{z}1})\} \\ &= \mathrm{Trace}\{E(\mathbf{X}_{\mathbf{z}1}^T E(\mathrm{diag}(\boldsymbol{\varepsilon}_{\mathbf{z}}^2) \mid \mathbf{X}_{\mathbf{z}1})\mathbf{X}_{\mathbf{z}1})\} \le c s_n n \end{aligned}$$

for some positive constant c. Together with $\lambda_n \rho'_{\lambda_n}(d_n) = O(n^{-1/2})$ and Markov's inequality, this yields

$$P\left(\max_{\mathbf{a} \in \mathcal{M}} Q_n(\mathbf{a}, \mathbf{0}; \mathbf{z}) - Q_n(\boldsymbol{\beta}^*) < 0\right) \ge P\left(\|\mathbf{X}_{\mathbf{z}1}^T \boldsymbol{\varepsilon}_{\mathbf{z}}\|_2^2 < 0.2c_3^2\tau^2 s_n n\right) \ge 1 - \frac{5}{\tilde{c}\tau^2}$$

for sufficiently large τ, where $\tilde{c}$ is a constant. This proves that $\hat{\boldsymbol{\beta}}_1$ lies in the interior of $\mathcal{M}$ such that $\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\|_2 = O_P(\sqrt{s_n/n})$.

(ii) Let $\mathbf{I}_{n21}(\boldsymbol{\beta}_1; \mathbf{z}) = -\partial^2 \tilde{L}_n(\boldsymbol{\beta}_1, \mathbf{0}; \mathbf{z})/\partial\boldsymbol{\beta}_2 \partial\boldsymbol{\beta}_1^T$. By the mean value of the theorem, we have

$$\|(n\lambda_n)^{-1}\mathbf{X}_{\mathbf{z}2}^T \boldsymbol{\varepsilon}_{\mathbf{z}}(\hat{\boldsymbol{\beta}}_1, \mathbf{0})\|_\infty \le \|(n\lambda_n)^{-1}\mathbf{X}_{2\mathbf{z}}^T \boldsymbol{\varepsilon}_{\mathbf{z}}(\boldsymbol{\beta}_1^*, \mathbf{0})\|_\infty + \|(\lambda_n)^{-1}\mathbf{I}_{n21}(\tilde{\boldsymbol{\beta}}_1, \mathbf{z})(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*)\|_\infty,$$

where $\tilde{\boldsymbol{\beta}}_1$ is between $\hat{\boldsymbol{\beta}}_1$ and $\boldsymbol{\beta}_1^*$.

Under Condition (A1), which is analogous to (S5.41) in the supplementary material of [18], there exist constants $\rho_1^* > 0$ such that

$$P(\max_{1 \le j \le p_n} \|\mathbf{X}_{\mathbf{z},j}\|_\infty \ge \rho_1^* \sqrt{\log(np_n)}) = O_P(n^{-1}p_n^{-1}),$$

where $\mathbf{X}_{\mathbf{z},j} = (x_{1z_1,j}, \ldots, x_{nz_n,j})$ and $j = 1, \ldots, p_n$. By the definition of the Orlicz norm, we obtain

$$\max_{1\le j\le p_n} E|X_{1z_1,j}|^2 \le \max_{1\le j\le p_n} \||X_{1z_1,j}|^2\|_{\psi_1} \le \max_{1\le j\le p_n} \|X_{1z_1,j}\|^2_{\psi_2},$$

where $\|\mathbf{X}_{1z_1,j}\|_{\psi_1} = \inf_{t>0}\{E\exp(t^{-1}|\mathbf{X}_{1z_1,j}|) \le 2\}$. Analogous to (E.22) in the supplementary material of [19], there exist constants $\rho_2^* > 0$ and $c > 0$ such that

$$P(\max_{1\le j\le p_n} \|\mathbf{X}_{\mathbf{z},j}\|_2^2 \ge \rho_2^* n) = O_P(p_n^{-1}\exp(-cn)).$$

Define $\delta_n^* = \{2c_2\}^{1/2}|\rho_2^*|^{1/2}\sqrt{n\log(np_n)}$. Similar to the proof of Proposition 4 in [17], we have

$$\begin{aligned} P(\|\mathbf{X}_{\mathbf{z}2}^T\boldsymbol{\varepsilon}_{\mathbf{z}}(\boldsymbol{\beta}_1^*, \mathbf{0})\|_\infty > \delta_n^* \mid \mathbf{X}_{\mathbf{z}}) &\le 2\sum_{j=1}^{p_n} \exp\left(-\frac{1}{2}\frac{\delta_n^{*2}}{c_2\|\mathbf{X}_{\mathbf{z},j}\|_2^2 + c_1\|\mathbf{X}_{\mathbf{z},j}\|_\infty\delta_n^*}\right) \\ &\le 2\sum_{j=1}^{p_n} \exp\left(-\frac{1}{2}\frac{\delta_n^{*2}}{c_2\rho_2^* n + c_1\rho_1^*\log^{1/2}(np_n)\delta_n^*}\right) \\ &\le 2/n \to 0. \end{aligned}$$

Combined with $\max\{\sqrt{\log(p_n)/n}, \sqrt{s_n/n}\} = o(\lambda_n)$, Condition (A3), and result (i), this yields

$$\|(n\lambda_n)^{-1}\mathbf{X}_{\mathbf{z}2}^T\boldsymbol{\varepsilon}_{\mathbf{z}}(\hat{\boldsymbol{\beta}}_1, \mathbf{0})\|_\infty = O_P(\lambda_n^{-1}\sqrt{\log(p_n)/n}) + O_P(\lambda_n^{-1}\sqrt{s_n/n}) \ll \rho'(0+)$$

for sufficiently large n.

(iii) By Conditions (A3) and (A5) and result (i), result (iii) holds.

Therefore, we have shown that $\hat{\boldsymbol{\beta}}^{(\mathbf{z})} = (\hat{\boldsymbol{\beta}}_1, \mathbf{0})$ is a strict local maximizer of (6) with probability going to one. □

Proof of Theorem 2. By the results in (8), we have

$$\hat{\beta}_d = \arg\min_{\beta_d}\{R^{-1}\sum_{r=1}^{R}(\tilde{\beta}_d^{(r)} - \beta_d)^2 + \lambda_n^{(d)}|\beta_d|\}$$

for $d = 1, \ldots, p_n$. The closed solution to the above penalized least squares estimation is

$$\hat{\beta}_d = \mathrm{sgn}(\bar{\beta}_d)\max\{|\bar{\beta}_d| - \lambda_n^{(d)}, 0\},$$

where $\bar{\beta}_d = R^{-1}\sum_{r=1}^{R}\tilde{\beta}_d^{(r)}$. Following Theorem 1, we have $P(\bar{\beta}_d = 0, d \in \mathcal{S}^c) \to 1$ and

$$|\bar{\beta}_d| \ge |\beta_d^*| - R^{-1}\sum_{r=1}^{R}|\tilde{\beta}_d^{(r)} - \beta_d^*| = |\beta_d^*| - O_P(\sqrt{s_n/n}) \ge d_n$$

for $d \in \mathcal{S}$. This and $\lambda_n^{(d)} = o(d_n)$ ensure that $P(\hat{\boldsymbol{\beta}}_{\mathcal{S}^c} = \mathbf{0}) \to 1$ as $n \to \infty$ and

$$\hat{\boldsymbol{\beta}}_{\mathcal{S}} = R^{-1}\sum_{r=1}^{R}\tilde{\boldsymbol{\beta}}_{\mathcal{S}}^{(r)},$$

which in combination with Theorem 1 yields

$$\|\hat{\boldsymbol{\beta}}_{\mathcal{S}} - \boldsymbol{\beta}_{\mathcal{S}}^*\|_2 = O_P(\sqrt{s_n/n}).$$

□

References

1. Wang, Y.G.; Fu, L.; Paul, S. *Analysis of Longitudinal Data with Examples*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2022.
2. Fitzmaurice, G.; Davidian, M.; Verbeke, G.; Molenberghs, G. (Eds.) *Longitudinal Data Analysis*, 1st ed.; Chapman and Hall/CRC: Boca Raton, FL, USA, 2008.
3. Liang, K.Y.; Zeger, S.L. Longitudinal data analysis using generalized linear models. *Biometrika* **1986**, *73*, 12–22. [CrossRef]
4. Hoffman, E.B.; Sen, P.K.; Weinberg, C.R. Within-cluster resampling. *Biometrika* **2001**, *88*, 1121–1134. [CrossRef]
5. Mancl, L.A.; Leroux, B.G. Efficiency of regression estimates for clustered data. *Biometrics* **1996**, *52*, 500–511. [CrossRef] [PubMed]
6. Wang, L.; Qu, A. Consistent model selection and datadriven smooth tests for longitudinal data in the estimating equations approach. *J. R. Stat. Soc. Ser.* **2009**, *71*, 177–190. [CrossRef]
7. Wang, L. GEE analysis of clustered binary data with diverging number of covariates. *Ann. Stat.* **2011**, *39*, 389–417. [CrossRef]
8. Wang, L.; Zhou, J.H.; Qu, A. Penalized Generalized Estimating Equations for High-Dimensional Longitudinal Data Analysis. *Biometrics* **2012**, *68*, 353–360. [CrossRef]
9. Fan, J.; Li, R. ariable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **2001**, *96*, 1348–1360. [CrossRef]
10. Fang, E.X.; Ning, Y.; Li, R. Test of significance for high-dimensional longitudinal data. *Ann. Stat.* **2020**, *48*, 2622–2645. [CrossRef]
11. Xia, L.; Shojaie, A. Statistical inference for high-dimensional generalized estimating equations. *arXiv* **2023**, arXiv:2207.11686.
12. Chiang, C.T.; Lee, K.Y. Efficient estimation methods for informative cluster size data. *Stat. Sin.* **2008**, *18*, 121–133.
13. Shen, W.C.; Chen, H.Y. Model selection for semiparametric marginal mean regression accounting for within-cluster subsampling variability and informative cluster size. *Biometrics* **2018**, *74*, 934–943. [CrossRef]
14. Wang, H.; Jiang, X.; Zhou, M.; Jiang, J. Variable selection for distributed sparse regression under memory constraints. *Commun. Math. Stat.* **2024**, *12*, 307–338. [CrossRef]
15. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. (Methodol.)* **1996**, *58*, 267–288. [CrossRef]
16. Zhang, C.H. Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.* **2010**, *38*, 894–942. [CrossRef]
17. Fan, J.; Lv, J. Non-concave penalized likelihood with NP-Dimensionality. *IEEE–Inf. Theory* **2011**, *57*, 5467–5484. [CrossRef] [PubMed]
18. Shi, C.; Song, R.; Chen, Z.; Li, R. Linear hypothesis testing for high dimensional generalized linear models. *Ann. Stat.* **2019**, *47*, 2671–2703. [CrossRef]
19. Shi, C.; Fan, A.; Song, R.; Lu, W. High-dimensional A-learning for optimal dynamic treatment regimes. *Ann. Stat.* **2018**, *46*, 925–957. [CrossRef]
20. Zhao, P.; Yu, B. On Model Selection Consistency of Lasso. *J. Mach. Learn. Res.* **2006**, *7*, 2541–2563.
21. Chen, J.; Chen, Z. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **2008**, *95*, 759–771. [CrossRef]
22. Spellman, P.T.; Sherlock, G.; Zhang, M.Q.; Iyer, V.R.; Anders, K.; Eisen, M.B.; Brown, P.O.; Botstein, D.; Futcher, B. Comprehensive Identification of Cell Cycle-Regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization. *Mol. Biol. Cell* **1998**, *9*, 3273–3297. [CrossRef]
23. Wang, L.; Chen, G.; Li, H. Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics* **2007**, *23*, 1486–1494. [CrossRef] [PubMed]
24. Lee, T.I.; Rinaldi, N.J.; Robert, F.; Odom, D.T.; Bar-Joseph, Z.; Gerber, G.K.; Hannett, N.M.; Harbison, C.T.; Thompson, C.M.; Simon, I.; et al. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* **2002**, *298*, 799–804. [CrossRef]
25. Loy, C.J.; Lydall, D.; Surana, U. Ndd1, a high-dosage suppressor ofcdc28-1n, is essential for expression of a subset of late-s-phase-specific genes in saccharomyces cerevisiae. *Mol. Cell. Biol.* **1999**, *19*, 3312–3327. [CrossRef] [PubMed]
26. Koch, C.; Moll, T.; Neuberg, M.; Ahorn, H.; Nasmyth, K. A role for the transcription factors mbp1 and swi4 in progression from g1 to s phase. *Science* **1993**, *261*, 1551–1557. [CrossRef]
27. Kumar, R.; Reynolds, D.M.; Shevchenko, A.; Shevchenko, A.; Goldstone, S.D.; Dalton, S. Forkhead transcription factors, fkh1p and fkh2p, collaborate with mcm1p to control transcription required for m-phase. *Curr. Biol.* **2000**, *10*, 896–906. [CrossRef] [PubMed]
28. Banerjee, N.; Zhang, M.Q. Identifying cooperativity among transcription factors controlling the cell cycle in yeast. *Nucleic Acids Res.* **2003**, *31*, 7024–7031. [CrossRef]
29. Spector, M.S.; Raff, A.; DeSilva, H.; Lee, K.; Osley, M.A. Hir1p and hir2p function as transcriptional corepressors to regulate histone gene transcription in the saccharomyces cerevisiae cell cycle. *Mol. Cell. Biol.* **1997**, *17*, 545–552. [CrossRef] [PubMed]
30. Koranda, M.; Schleiffer, A.; Endler, L.; Ammerer, G. Forkhead-like transcription factors recruit ndd1 to the chromatin of g2/m-specific promoters. *Nature* **2000**, *406*, 94–98. [CrossRef]
31. Rao, A.R.; Pellegrini, M. Regulation of the yeast metabolic cycle by transcription factors with periodic activities. *BMC Syst. Biol.* **2011**, *5*, 160. [CrossRef]
32. Carrillo, E.; Ben-Ari, G.; Wildenhain, J.; Tyers, M.; Grammentz, D.; Lee, T.A. Characterizing the roles of Met31 and Met32 in coordinating Met4-activated transcription in the absence of Met30. *Mol. Biol. Cell.* **2012**, *23*, 1928–1942. [CrossRef]

33. Blank, H.M.; Gajjar, S.; Belyanin, A.; Polymenis, M. Sulfur metabolism actively promotes initiation of cell division in yeast. *PLoS ONE* **2009**, *4*, e8018. [CrossRef]
34. Pic, A.; Lim, F.L.; Ross, S.J.; Veal, E.A.; Johnson, A.L.; Sultan, M.R.; West, A.G.; Johnston, L.H.; Sharrocks, A.D.; Morgan, B.A. The forkhead protein Fkh2 is a component of the yeast cell cycle transcription factor SFF. *EMBO J.* **2000**, *19*, 3750–3761. [CrossRef] [PubMed]
35. Foord, R.; Taylor, I.A.; Sedgwick, S.G.; Smerdon, S.J. X-ray structural analysis of the yeast cell cycle regulator Swi6 reveals variations of the ankyrin fold and has implications for Swi6 function. *Nat. Struct. Biol.* **1999**, *6*, 157–165.
36. Oehlen, L.J. Ste12 and Mcm1 regulate cell cycle-dependent transcription of FAR1. *Mol. Cell. Biol.* **1996**, *16*, 2830–2837. [CrossRef] [PubMed]
37. Bouquin, N.; Johnson, A.L.; Morgan, B.A.; Johnston, L.H. Association of the cell cycle transcription factor Mbp1 with the Skn7 response regulator in budding yeast. *Mol. Biol. Cell.* **1999**, *10*, 3389–3400. [CrossRef] [PubMed]
38. Osley, M.A.; Lycan, D. Trans-acting regulatory mutations that alter transcription of Saccharomyces cerevisiae histone genes. *Mol. Cell. Biol.* **1987**, *7*, 4204–4210.
39. Moser, S.C.; Bensaddek, D.; Ortmann, B.; Maure, J.F.; Mudie, S.; Blow, J.J.; Lamond, A.I.; Swedlow, J.R.; Rocha, S. PHD1 links cell-cycle progression to oxygen sensing through hydroxylation of the centrosomal protein Cep192. *Dev. Cell.* **2013**, *26*, 381–392. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.