# Estimation and variable selection in multicollinear regression model using broken adaptive Liu-type estimator

Adewale F. Lukman[1*†], Essoham Ali[2,3†] and Emmanuel T. Adewuyi[4†]

[1]Department of Mathematics and Statistics, University of North Dakota, Grand Forks, North Dakota, 58202 USA.
[1]Institut de Mathématiques Appliquées, UCO, 49000, Angers, France.
[2]Univ Bretagne Sud, CNRS UMR 6205, LMBA, Vannes, France.
[4]Department of Medical Statistics, London School of Hygiene and Tropical Medicine, Keppel St., London, WC1E 7HT UK.

*Corresponding author(s). E-mail(s): adewale.lukman@und.edu;
Contributing authors: essoham.ali@univ-ubs.fr ; emmanuel.adewuyi@lshtm.ac.uk ;
†These authors contributed equally to this work.

**Abstract**

This paper introduces the Broken Adaptive Liu-Type (BALT) estimator, a novel penalized regression approach designed to achieve accurate parameter estimation and effective variable selection in high-dimensional settings. By integrating adaptive shrinkage with a broken weighting mechanism, BALT enables differential regularization across components of the parameter space, facilitating both sparsity and stability. The estimator is derived from the Liu-type family and incorporates a flexible structure that adaptively targets relevant predictors while shrinking negligible coefficients toward zero. We establish the large-sample properties of BALT, including its oracle property and grouping effect, under general non-orthogonal conditions. Extensive simulation studies and real-data applications, including prostate cancer, electricity consumption, and riboflavin gene expression, demonstrate that BALT consistently achieves lower prediction error and more parsimonious models than existing methods such as Lasso, Elastic Net, Ridge, and Broken Adaptive Ridge. These results highlight BALT as a powerful and interpretable tool for sparse modeling in complex, high-dimensional regression problems.

**Keywords:** Sparse estimation; Variable selection; Biased regression; Liu-type estimator; Oracle property; High-dimensional inference.

## 1 Introduction

Linear regression models are fundamental tools in statistical analysis, widely applied across diverse fields such as social sciences, biology, economics, and engineering [30]. The primary objective of these models is to establish a linear relationship between a dependent variable $y$ and a set of explanatory variables represented by the design matrix $X \in \mathbb{R}^{n\times p}$, where $n$ denotes the number of observations and $p$ the number of variables.

However, when certain explanatory variables are highly correlated, a phenomenon known as multicollinearity the matrix $X^\top X$ becomes singular or nearly singular. This non-invertibility or ill-conditioning undermines the estimation of coefficients using the Ordinary Least Squares (OLS) method. Specifically, a non-invertible $X^\top X$ matrix prevents the unique calculation of regression coefficients, leading to model instability and unreliable estimates.

Biased estimators have been developed to mitigate these effects. Ridge regression, introduced by Hoerl and Kennard [19], was among the first alternatives to incorporate an $\ell_2$ penalty to stabilize estimates. While it reduces variance, it does not ensure strict variable selection. The Liu estimator [23], an extension of Ridge regression, integrates an adjustable bias structure that enhances stability in the presence of multicollinearity. Nonetheless, this approach does not guarantee sparsity in high-dimensional environments, limiting its effectiveness for interpretation and prediction.

Simultaneous variable selection and parameter estimation play a crucial role in statistical modeling and its broad range of applications. A straightforward and intuitive strategy for variable selection involves the use of $\ell_0$-penalized regression, which imposes a penalty based on the number of selected variables. This approach is closely linked to traditional model selection criteria such as Akaike Information Criterion (AIC) [1], and Bayesian Information Criterion (BIC) [3]. The $\ell_0$ penalty has been shown to possess desirable theoretical properties, offering optimal performance in both variable selection and parameter estimation. However, the associated optimization problem is inherently nonconvex and requires an exhaustive search over all possible subsets of variables a task that is NP-hard and computationally prohibitive even for moderately sized datasets. Furthermore, solutions obtained via this method may exhibit instability in variable selection.

To address these challenges, the $\ell_1$-penalized regression method, known as the Lasso, has emerged as a widely adopted alternative. Lasso enjoys consistency in variable selection, although it falls short in delivering consistent parameter estimation. Over the past two decades, significant research efforts have been directed towards refining the Lasso through various extensions of the $\ell_1$ penalty [2, 5–7, 9, 32, 33]. These advanced techniques aim to achieve both selection consistency and estimation consistency, thereby enhancing the reliability and interpretability of statistical models.

In this context, the Broken Adaptive Ridge (BAR) method has been recently introduced [4]. It combines the strengths of $\ell_0$ penalties (promoting sparsity) and $\ell_2$ penalties (stability against multicollinearity) by adaptively adjusting penalty weights across different model components. This "broken" structure allows for the strict nullification of certain coefficients while stabilizing others, offering a more flexible and effective solution.

Building upon this, our work proposes a new class of biased estimators: the Broken Adaptive Liu-Type Estimator (BALT). Inspired by Liu-type estimators, BALT introduces two distinct bias parameters within a broken adaptive structure, enabling fine regulation of multicollinearity and automatic variable selection. The method relies on an orthogonal decomposition of the design matrix, separating components associated with large and small eigenvalues the latter often being the source of instability in the presence of multicollinearity.

The BALT approach aims to simultaneously improve parameter estimation and variable selection in high-dimensional contexts. It benefits from a rigorous theoretical framework, demonstrating asymptotic properties such as consistency, relative efficiency, and the oracle property. Extensive simulations, comparing BALT to existing methods like Ridge, Lasso, BAR, and Elastic Net, highlight the superiority of our approach in scenarios characterized by strong multicollinearity and a small number of relevant variables.

In summary, this work contributes to the ongoing efforts to enhance linear models' performance in high-dimensional settings. The BALT estimator, with its innovative architecture based on adaptive and differentiated weighting, represents a significant advancement in variable selection, estimation stability, and model interpretability in complex contexts. In Section 2, we review the estimation and variable selection methods. In Section 3, under certain regularity conditions, we state our main large-sample results for BALT, namely the oracle and grouping properties for the general non-orthogonal case. Section 4 illustrates the simulation results. Section 5 presents applications involving prostate cancer data, electricity consumption data, and data on the composition of bituminous binder and surface free energy. Concluding remarks are given in Section 6. Technical proofs are postponed to an Appendix.

## 2 Estimation and variable selection methods

The linear regression model (LRM) expresses the response variable $y$ as a linear combination of one or more predictors $z$. The general LRM in matrix form is given by:

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{y} \in \mathbb{R}^n$ is the vector of responses, $\mathbf{Z} \in \mathbb{R}^{n\times p}$ is the design matrix of predictors, $\boldsymbol{\theta} \in \mathbb{R}^p$ is the vector of the true regression coefficient and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is the vector of error terms with mean zero and variance $\sigma^2\mathbf{I}_n$. The design matrix $\mathbf{Z}$ is standardized so that $\sum_{i=1}^n z_{ij}/n = 0$, $\sum_{i=1}^n z_{ij}^2/n = 1$ for $j = 1, 2, \ldots, p$ and the response vector $\mathbf{y}$ is centered to have mean zero $\sum_{i=1}^n y_i/n = 0$. This allowed us to fit a regression model without the intercept term. The Least squares estimator (LSE) is obtained by solving the following optimization problem

$$\hat{\boldsymbol{\theta}}^{LS} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p}\{\|\mathbf{y} - \mathbf{Z}\boldsymbol{\theta}\|_2^2\} = (Z^TZ)^{-1}Z^Ty. \tag{2}$$

The least squares estimator (LSE) is an unbiased estimator with minimum variance in the absence of multicollinearity. However, when multicollinearity is present, the variance of the LSE increases significantly, leading to large coefficient estimates [8]. In addition, in high-dimensional settings, where the number of predictors exceeds the number of observations, the design matrix $\mathbf{Z}$ is not of full rank, resulting in the absence of a unique solution [10, 11].

Penalized regression methods serve as effective alternatives to LSE by mitigating issues related to large coefficient estimates and multicollinearity. These methods impose constraints on the magnitude of the coefficient estimates of $\boldsymbol{\theta}$ in model (1), leading to more stable solutions. The penalized coefficient estimates are obtained by solving the following optimization problem:

$$\hat{\boldsymbol{\theta}}(\lambda, \nu) = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p}\{\|\mathbf{y} - \mathbf{Z}\boldsymbol{\theta}\|_2^2 + \lambda p_\nu(\boldsymbol{\theta})\}, \tag{3}$$

where $p_\nu(\cdot)$ is the penalization function that describes the form of penalization and $\lambda$ is a tuning parameter that controls the magnitude of the coefficient estimates of $\boldsymbol{\theta}$. The ridge estimator represents a form of regularization known as $\ell$-norm penalization, which introduces a constraint on the magnitude of the regression coefficients to improve estimation stability and mitigate multicollinearity (Hoerl and Kennard, 1970). It is obtained by solving the following optimization problem

$$\hat{\boldsymbol{\theta}}^{(r)} = \arg\min_{\boldsymbol{\theta}}\left\{\|\mathbf{y} - \mathbf{Z}\boldsymbol{\theta}\|^2 + \lambda_n\sum_{j=1}^{p_n}\theta_j^2\right\} = (\mathbf{Z}^T\mathbf{Z} + \lambda_n\mathbf{I})^{-1}\mathbf{Z}^T\mathbf{y}, \tag{4}$$

where $\lambda_n \geq 0$ is a regularization parameter. Ridge regression effectively addresses multicollinearity and provides unique coefficient estimates in high-dimensional settings. However, it does not induce sparsity, as none of its estimates are shrunk to zero, making interpretation challenging. To overcome this limitation, Dai et al. [4] introduced the broken adaptive ridge (BAR) regression, a sparse extension of ridge regression, with the objective function given by:

$$\hat{\boldsymbol{\theta}}^{BAR} = \arg\min_{\boldsymbol{\theta}}\left\{\|\mathbf{y} - \mathbf{Z}\boldsymbol{\theta}\|^2 + \lambda_n\sum_{j=1}^{p_n}\frac{\theta_j^2}{\hat{\theta}_j^2}\right\} = (\mathbf{Z}^T\mathbf{Z} + \lambda_n\mathbf{D}(\hat{\boldsymbol{\theta}}))^{-1}\mathbf{Z}^T\mathbf{y}, \tag{5}$$

where $\mathbf{D}(\hat{\boldsymbol{\theta}}) = \text{diag}(\hat{\theta}_1^{-2}, \ldots, \hat{\theta}_{p_n}^{-2})$.

Despite its advantages, the formulation in equation (5) may suffer from numerical instability, as division by small $\hat{\theta}_j^2$ values can lead to excessive penalization. A common remedy is to introduce a small perturbation $\delta > 0$ to $\mathbf{D}(\hat{\boldsymbol{\theta}})$, ensuring numerical stability. Alternatively, equation (5) can be reformulated as:

$$\hat{\boldsymbol{\theta}}^{BAR} = \boldsymbol{\Gamma}(\hat{\boldsymbol{\theta}})\mathbf{Z}^T(\mathbf{Z}^T(\hat{\boldsymbol{\theta}})\mathbf{Z}(\hat{\boldsymbol{\theta}}) + \lambda_n\mathbf{I}_n)^{-1}\mathbf{Z}^T(\hat{\boldsymbol{\theta}})\mathbf{y}, \tag{6}$$

where $\boldsymbol{\Gamma}(\hat{\boldsymbol{\theta}}) = \text{diag}(\hat{\theta}_1, \ldots, \hat{\theta}_{p_n})$ and $\mathbf{Z}(\hat{\boldsymbol{\theta}}) = \mathbf{Z}\boldsymbol{\Gamma}(\hat{\boldsymbol{\theta}})$.

This alternative formulation preserves numerical stability without requiring an explicit perturbation term, making it a robust approach for high-dimensional regression problems. Liu [23], Ozkale and Kaciranlar [31] and Liu [24] developed the Liu-type estimator (LTE) as a competitive alternative to ridge regression. In this study, the LTE is defined as follows:

$$\hat{\boldsymbol{\theta}}^{LT} = \arg\min_{\boldsymbol{\theta}}\left\{\|\mathbf{y} - \mathbf{Z}\boldsymbol{\theta}\|^2 + \lambda_n\|\theta - (-d)\boldsymbol{\theta}^{LS}\|^2\right\} = (\mathbf{Z}^T\mathbf{Z} + \lambda_nI_n)^{-1}(\mathbf{Z}^T\mathbf{Z} - \lambda d)\theta^{LS} \tag{7}$$

LTE exhibit similar performance to ridge regression, but cannot induce sparsity and fails to provide a unique solution in high-dimensional settings. Therefore, in this study, we propose the Broken

Adaptive Liu-Type Estimator (BALT), which is designed to induce sparsity and remain effective in high-dimensional contexts. Hence, the penalization functions is defined as follows: For any integer $t \geq 1$, define

$$\hat{\theta}^t = f(\hat{\theta}^{t-1}), \tag{8}$$

where

$$f(\tilde{\boldsymbol{\theta}}) = \arg\min_{\boldsymbol{\theta}} \left\{ \|\mathbf{y} - \mathbf{Z}\boldsymbol{\theta}\|^2 + \frac{\lambda\|\theta - (-d)\tilde{\theta}_j \boldsymbol{\theta}^{LS}\|^2}{\tilde{\theta}_j^2} \right\} = (\mathbf{Z}^T\mathbf{Z} + \lambda_n \tilde{\theta}_j^{-2} I_n)^{-1}(\mathbf{Z}^T\mathbf{Z} - \lambda d\tilde{\theta}_j^{-1})\theta^{LS}. \tag{9}$$

The broken adaptive Liu-type (BALT) estimator is defined as

$$\hat{\boldsymbol{\theta}}^* = \lim_{t\to\infty} \hat{\boldsymbol{\theta}}^{(t)}.$$

To ensure numerical stability, we rewrite (9) as follows:

$$f(\tilde{\boldsymbol{\theta}}) = (\mathbf{Z}^T D(\tilde{\theta})\mathbf{Z} + \lambda_n I_n)^{-1}(\mathbf{Z}^T D(\tilde{\theta})\mathbf{Z} - \lambda_n d F(\tilde{\theta}))\theta^{LS}, \tag{10}$$

where $F(\tilde{\theta}) = diag(\tilde{\theta}_j)$ and $D(\tilde{\theta}) = diag(\tilde{\theta}_j^2)$.

Following Dai et al. [4], we assume that $Z^T Z/n = I_n$. Then for each $j \in \{1, ..., n\}$, the jth component of $f(\tilde{\boldsymbol{\theta}})$ defined by (9) is as follows:

$f_j(\tilde{\boldsymbol{\theta}}_j) = \frac{(\tilde{\theta}_j^2 - \frac{d\lambda_n \tilde{\theta}_j}{n})}{(\tilde{\theta}_j^2 + \frac{\lambda_n}{n})}\hat{\theta}_j^{LS}$ where $\theta_j^{LS} = n^{-1}Z^T y$.

**Remark 1** (Orthogonal Case)**.** *It is important to examine the orthogonal case, which admits a closed-form expression for the BALT estimator. Without loss of generality, assume that* $\theta_i^{\mathrm{LS}} > 0$ *and* $0 \leq \hat{\theta}_j^{\lambda d} \leq \hat{\theta}_j^{LS}$ *for every integer* $k$*. Also, note that the map*

$$z \;\mapsto\; f(z) = \frac{(z^2 - \frac{\lambda_n d}{n} z)\hat{\theta}_j^{LS}}{(z^2 + \lambda_n/n)}$$

*is increasing in* $z$ *on* $(0, \infty)$*. If we set* $\phi(z) = f(z) - z$ *and fixed the value of* $n, \lambda$ *and* $d$*, we have the following observations:*

1. *If* $(\hat{\theta}_j^{LS})^2 < \frac{4\lambda_n}{n}(1 + d\hat{\theta}_j^{LS})$*, then* $\phi(z) < 0$*. Therefore,* $\hat{\theta}_j^{(t+1)} < \hat{\theta}_j^{(t)}$ *for all* $t$ *and* $\hat{\theta}_j^* = 0$*.*
2. *If* $(\hat{\theta}_j^{LS})^2 = \frac{4\lambda_n}{n}(1 + d\hat{\theta}_j^{LS})$*, then* $\phi(z) \leq 0$*. Therefore,* $\hat{\theta}_j^{(t+1)} \leq \hat{\theta}_j^{(t)}$ *for all* $t$*. Then,*

$$\hat{\theta}_j^* = \begin{cases} 0, & \text{if } \hat{\theta}_j^{(0)} < \frac{\hat{\theta}_j^{LS}}{2}, \\ \frac{\hat{\theta}_j^{LS}}{2}, & \text{otherwise.} \end{cases}$$

3. *If* $(\hat{\theta}_j^{LS})^2 \geq 4\lambda_n/n(1 + d\theta_j^{LS})$*, then* $\phi(x) \geq 0$ *when* $x \in [x_1, x_2]$*, where*

$$x_1 = \frac{\hat{\theta}_j^{LS}}{2} - \sqrt{\frac{(\hat{\theta}_j^{OLS})^2}{4} - \frac{\lambda_n}{n}(1 + d\hat{\theta}_j^{LS})}$$

*and*

$$x_2 = \frac{\hat{\theta}_j^{LS}}{2} + \sqrt{\frac{(\hat{\theta}_j^{OLS})^2}{4} - \frac{\lambda_n}{n}(1 + d\hat{\theta}_j^{LS})}$$

*and* $\phi(x) < 0$ *otherwise. As a result,*

$$\hat{\theta}_j^* = \begin{cases} 0, & \text{if } \hat{\theta}_j^{(0)} < z_1, \\ \frac{\hat{\theta}_j^{LS}}{2} + \sqrt{\frac{(\hat{\theta}_j^{OLS})^2}{4} - \frac{\lambda}{n}(1 + d\hat{\theta}_j^{LS})}, & \text{otherwise.} \end{cases}$$

*The BALT estimator typically converges to a unique solution, with its convergence properties being influenced by the chosen initial value. In this work, we employ the ridge estimator as the initial value. As a result, the BALT algorithm converges for the ith component to the following limit:*

$$\hat{\theta}_j^* = \begin{cases} 0, & \text{if } \hat{\theta}_j^{(LS)} \in (-2\sqrt{\lambda_n/n(1+d\hat{\theta}_j^{LS})}), \\ \frac{\hat{\theta}_j^{LS}}{2} + \sqrt{\frac{(\hat{\theta}_j^{OLS})^2}{4} - \frac{\lambda_n}{n}(1+d\hat{\theta}_j^{LS})}, & \text{if } \hat{\theta}_j^{(LS)} \notin (-2\sqrt{\lambda_n/n(1+d\hat{\theta}_j^{LS})}). \end{cases} \tag{11}$$

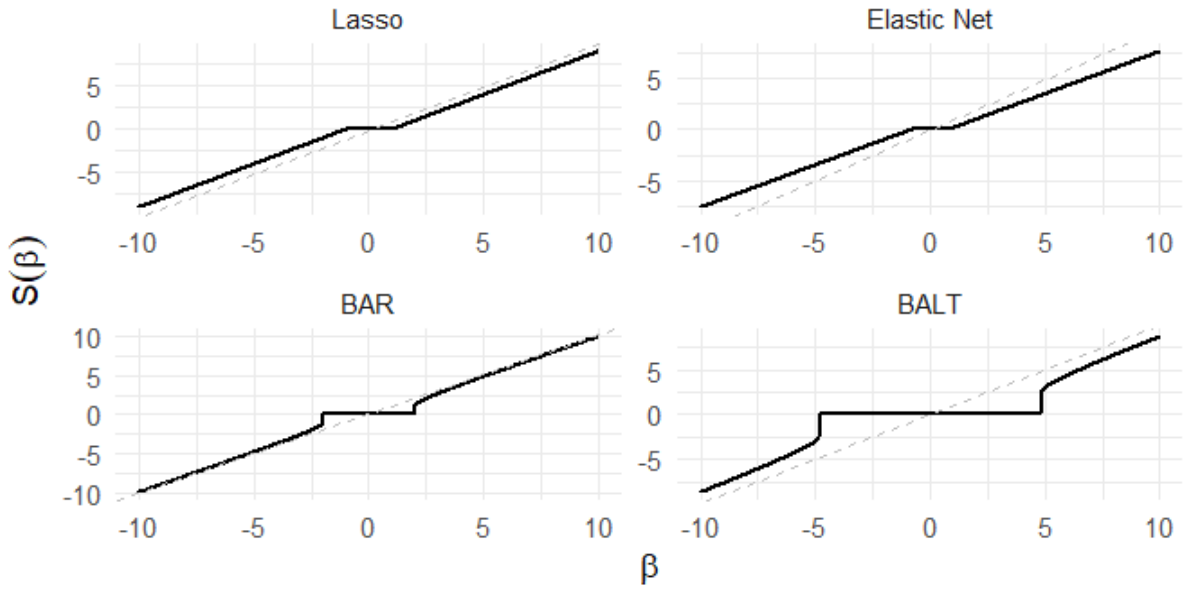


**Figure 1**: Thresholding functions of BAR, BALT, LASSO and Enet.

Figure 1 illustrates the thresholding functions for BALT (as defined in (11)), BAR, Lasso, and Elastic Net. The plots reveal distinct shrinkage behaviors inherent to each regularization method. In particular, the Lasso and Elastic Net functions exhibit smooth, gradual transitions around the threshold, reflecting their continuous soft-thresholding operations. In contrast, the BAR and BALT methods display more abrupt, piecewise changes, inducing exact zeros for coefficients below a specified threshold. Notably, while the BALT function behaves similarly to BAR in enforcing sparsity, its formulation includes an additional tuning parameter that allows for a more flexible adjustment of the shrinkage intensity. This added flexibility can be advantageous in accommodating varying degrees of sparsity and improving model performance in high-dimensional settings.

# 3 Large-sample properties of BALT

## 3.1 Oracle Property

In this section, we investigate the oracle properties of the BALT estimator under a general framework without imposing the condition that the design matrix must be orthogonal.

Let $\theta_0 = (\theta_{01}, \ldots, \theta_{0p_n})^\top$ represent the true parameter vector, where the dimensionality $p_n$ increases with the sample size but satisfies $p_n < n$. For clarity, decompose $\theta_0$ as $\theta_0 = (\theta_{01}^\top, \theta_{02}^\top)^\top$, where $\theta_{01} \in \mathbb{R}^{q_n}$ contains the nonzero elements and $\theta_{02} \in \mathbb{R}^{p_n - q_n}$ consists of zero coefficients. We assume $\theta_{01} \neq 0$ and $\theta_{02} = 0$. Let $\hat{\theta}^* = (\hat{\theta}_1^{*\top}, \hat{\theta}_2^{*\top})^\top$ denote the BALT estimator of $\theta_0$. Define $\mathbf{Z}_1 = (\mathbf{z}_1, \ldots, \mathbf{z}_{q_n})$, $\Sigma_{n1} = \mathbf{Z}_1^\top \mathbf{Z}_1/n$, and $\Sigma_n = \mathbf{Z}^\top \mathbf{Z}/n$.

The following assumptions are necessary to derive the asymptotic properties of the BALT estimator.

(C1) The random errors $\varepsilon_1, \ldots, \varepsilon_n$ are independent and identically distributed with mean zero and finite variance $0 < \sigma^2 < \infty$.

(C2) There exists a constant $C > 1$ such that for all $n$, the eigenvalues of $\Sigma_n$ are bounded: $0 < 1/C < \lambda_{\min}(\Sigma_n) \leq \lambda_{\max}(\Sigma_n) < C < \infty$.

(C3) Define $b_{0n} = \min_{1 \leq j \leq q_n} |\theta_{0j}|$ and $b_{1n} = \max_{1 \leq j \leq q_n} |\theta_{0j}|$. Then as $n \to \infty$, we require that $p_n q_n / \sqrt{n} \to 0$, $\xi_n b_{1n} / \sqrt{n} \to 0$, $(p_n/n)^{1/2}/b_{0n} \to 0$, and $\lambda_n b_{1n} (q_n/n)^{1/2} / b_{0n}^2 \to 0$.

**Theorem 1** (Oracle Property)**.** *Suppose that assumptions (C1)–(C3) are satisfied. For any vector* $\mathbf{a}_n \in \mathbb{R}^{q_n}$ *with* $\|\mathbf{b}_n\| \leq 1$*, define the scaling quantity as* $s_n^2 = \sigma^2 \mathbf{a}_n^\top \Sigma_{n1}^{-1} \mathbf{a}_n$*. Consider the function*

$$f(\boldsymbol{\alpha}) = \left(\mathbf{Z}_1^T \mathbf{Z}_1 + \lambda_n D_1(\alpha)\right)^{-1} \left(\mathbf{Z}_1^\top \mathbf{Z}_1 - \lambda_n d \, \mathrm{diag}(\tilde{\theta}_1^{-1}, \ldots, \tilde{\theta}_{q_n}^{-1})\right) (\boldsymbol{Z}_1^T \boldsymbol{Z}_1)^{-1} \boldsymbol{Z}_1^T \boldsymbol{y},$$

*where* $d$ *is a tuning parameter. Then, the fixed point of* $f(\tilde{\boldsymbol{\theta}})$ *exists and is unique.*

*Moreover, with probability approaching one as* $n \to \infty$*:*

*(i) The BALT estimator* $\hat{\boldsymbol{\theta}}^* = \left(\hat{\boldsymbol{\theta}}_1^{*\top}, \hat{\boldsymbol{\theta}}_2^{*\top}\right)^\top$ *exists and is unique, where* $\hat{\boldsymbol{\theta}}_2^* = 0$ *and* $\hat{\boldsymbol{\theta}}_1^*$ *is the unique fixed point of* $f(\boldsymbol{\alpha})$*;*

*(ii) The scaled error converges in distribution:*

$$\sqrt{n} s_n^{-1} \mathbf{a}_n^\top (\hat{\boldsymbol{\theta}}_1^* - \boldsymbol{\theta}_{01}) \xrightarrow{d} \mathcal{N}(0, 1).$$

**Theorem 2** (Grouping Property of BALT)**.** *Assume that the columns of the design matrix* $\mathbf{Z}$ *are standardized such that* $\sum_{k=1}^n z_{ik} = 0$ *and* $\|\mathbf{z}_i\|_2^2 = 1$ *for all* $i \in \{1, \ldots, p\}$*. Let* $\hat{\boldsymbol{\theta}}^*$ *be the BALT estimator given by Equation (9). Then, with probability tending to 1, for any* $i < j$*,*

$$|\hat{\theta}_i^{*-1} - \hat{\theta}_j^{*-1}| \leq \frac{1}{\lambda_n} \|\mathbf{y}\| \sqrt{2(1 - \rho_{ij})} \cdot \left| \frac{\tilde{\theta}_j}{1 + d\tilde{\theta}_j} - \frac{\tilde{\theta}_i}{1 + d\tilde{\theta}_i} \right|, \tag{12}$$

*provided that* $\hat{\theta}_i^* \cdot \hat{\theta}_j^* \neq 0$*, where* $\rho_{ij} = \mathbf{z}_i^T \mathbf{z}_j$ *is the sample correlation between* $\mathbf{z}_i$ *and* $\mathbf{z}_j$*, and* $\tilde{\theta}_i$*,* $\tilde{\theta}_j$ *are the reweighting parameters associated with the BALT penalty.*

# 4 Simulation

In this section, we present a simulation study to evaluate the predictive performance and variable selection capabilities of BALT in comparison with BAR, Lasso, Elastic Net, Ridge, and Adaptive Lasso. The simulated data follow the model

$$y = Z\boldsymbol{\theta} + \sigma \, \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, 1), \tag{13}$$

where $y$ is the response vector, $Z$ is the design matrix, $\boldsymbol{\theta}$ is the vector of regression coefficients, and $\sigma$ denotes the noise level.

We consider five distinct scenarios, mostly adapted from the BAR study [4]. Scenario 3 is specifically designed to explore grouped variable behavior, highlighting the potential advantages of BALT in structured settings.

To evaluate each method, we use $k$-fold cross-validation to approximate the train/validation/test split. For each replicate, data are partitioned into $k$ folds, where each fold serves once as validation while the remaining folds form the training set. Regularization parameters are selected via grid search to minimize the mean squared error (MSE) on the validation set.

Once optimal hyperparameters are determined, we re-fit the models on the full dataset and assess their performance using several criteria: prediction accuracy (MSE and mean absolute error (MAE)) and variable selection quality via the false positive rate (FPR) and false negative rate (FNR). A low FPR indicates few irrelevant variables are selected, while a low FNR implies strong retention of relevant variables.

We aggregate results over 50 replicates to ensure robustness across all methods and scenarios. The details for each scenario are outlined below:

- **Scenario 1:** We generate 50 datasets under two conditions: one with $p = 10$ and another with $p = 50$ predictors, both with $n = 100$ observations. The true coefficient vector is defined as $\boldsymbol{\beta} = (2, -3, 0, 0, 4, 0, 0, \ldots, 0)$, and the error term follows $\sigma = 1$. To examine the influence of correlation, we fix the pairwise correlation structure via $\Sigma_{kj} = r^{|k-j|}, \quad r \in \{0.5, 0.7, 0.9\}$. Results are presented in Table 1.
- **Scenario 2:** This scenario mirrors Scenario 1 but introduces additional weak signals: $\boldsymbol{\beta} = (2, -3, 0, 0, 4, 0.2, -0.3, 0, 0, 0, 0.4, 0, \ldots, 0)$. Results are shown in Table 2.
- **Scenario 3:** This setting introduces grouped variables. Let $p = 9$, and define the true coefficient vector as $\boldsymbol{\beta}_0 = (2, 2, 0.4, 0.4, 0, 0, 0, 0, 0)$, with responses generated by $\boldsymbol{y} = \boldsymbol{Z\theta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, 0.2^2 \boldsymbol{I}_n)$.
  The design matrix is constructed as follows:

$$\begin{aligned} \boldsymbol{z}_i &= \boldsymbol{x}_1 + \boldsymbol{e}_i, \quad \boldsymbol{x}_1 \sim \mathcal{N}(\mathbf{0}, 30^2 \boldsymbol{I}_n), \quad i \in \{1, 2, 3\}, \\ \boldsymbol{z}_i &= \boldsymbol{x}_2 + \boldsymbol{e}_i, \quad \boldsymbol{x}_2 \sim \mathcal{N}(\mathbf{0}, 30^2 \boldsymbol{I}_n), \quad i \in \{4, 5, 6\}, \\ \boldsymbol{z}_i &\sim \mathcal{N}(\mathbf{0}, 30^2 \boldsymbol{I}_n), \quad i \in \{7, 8, 9\}, \end{aligned}$$

  where $\boldsymbol{e}_i \sim \mathcal{N}(0, 0.1^2 \boldsymbol{I}_n)$. Predictors $\boldsymbol{z}_1$ to $\boldsymbol{z}_3$ and $\boldsymbol{z}_4$ to $\boldsymbol{z}_6$ form two latent groups, with within-group correlation of 1 and between-group correlation of 0.1. Sample size is $n = 200$. Results are summarized in Table 3.
- **Scenario 4 (High-Dimensional):** We generate 50 datasets with $n = 50$ and $p = 100$. The true coefficient vector is $\boldsymbol{\beta}' = (1, 1, 1, 1, 1, 0, \ldots, 0)$, with $\sigma = 3$ [12]. Covariance is defined as $\Sigma_{kj} = r^{|k-j|}$ for $r = 0.7$ and 0.9. Results appear in Table 4.
- **Scenario 5:** Another high-dimensional case with $n = 50$, $p = 100$, and true coefficient vector $\boldsymbol{\beta} = (2, -3, 0, 0, 4, 0.2, -0.3, 0, 0, 0, 0.4, 0, \ldots, 0)$, drawn from $\mathcal{N}(0, \Sigma)$ with $\Sigma_{kj} = 0.5^{|k-j|}$. Results are in Table 5.

*Results:* In Scenario 1, BALT consistently outperformed other methods, mostly in terms of MSE, while promoting model sparsity. BAR achieved the best FPR and FNR scores, with BALT a close second. This performance remained stable across different values of $p$ and $r$. Scenario 2 showed a similar pattern, despite the increased number of nonzero coefficients. While Lasso, Elastic Net, and Ridge improved slightly in FPR and FNR, particularly at higher $r$ values, BAR and BALT remained competitive in MSE and MAE. Their alternating performance across metrics reflects structural similarities in their algorithms. Notably, BALT consistently attained the smallest FPR, demonstrating exceptional parsimony without sacrificing estimation precision. In Scenario 3, BAR achieved the sparsest models (FPR = 0.1133), followed by BALT (FPR = 0.1400), though both exhibited relatively high FNRs (0.6267 and 0.6700, respectively). Ridge regression minimized FNR (0.0000) but included all predictors (FPR = 1.0000). Lasso and Elastic Net showed balanced FPR and FNR values but were still outperformed in terms of predictive accuracy by BAR and BALT. Scenario 4 results (Table 4) confirmed BALT's robustness in high-dimensional settings. It attained the lowest MSE (7.9453 for $r = 0.7$, 8.4418 for $r = 0.9$) and very low FPRs (0.0162, 0.0095). BAR followed closely in MSE and FPR but had slightly higher FNRs. Lasso and Enet performed moderately, while Ridge again achieved FNR = 0.0 with FPR = 1.0, reflecting severe overselection. In Scenario 5, BALT maintained strong performance across all metrics. BAR slightly surpassed BALT in MAE and FPR, while Lasso and Enet achieved lower FNRs. Overall, BAR and BALT demonstrated the most favorable trade-offs in both sparse and grouped settings.

**Table 1**: Performance comparison for different methods under varying $r$ and $p$ for scenario 1.

| | | $p = 10$ | | | | $p = 50$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| $r$ | Method | MSE | MAE | FPR | FNR | MSE | MAE | FPR | FNR |
| | Lasso | 1.1158 | 0.0726 | 0.5214 | 0 | 1.1415 | 0.0272 | 0.2294 | 0 |
| | Ridge | 1.3085 | 0.2168 | 1 | 0 | 1.0047 | 0.1615 | 1 | 0 |
| $r = 0.5$ | ElasticNet | 1.0952 | 0.0846 | 0.6900 | 0 | 1.1308 | 0.0398 | 0.3589 | 0 |
| | BAR | 0.9534 | 0.0372 | 0.0529 | 0 | 0.9336 | 0.0079 | 0.0085 | 0 |
| | BALT | 0.9524 | 0.0375 | 0.0557 | 0 | 0.9222 | 0.0096 | 0.0153 | 0 |
| | Lasso | 1.1063 | 0.0914 | 0.5600 | 0 | 1.1251 | 0.0341 | 0.2602 | 0 |
| | Ridge | 1.4495 | 0.3308 | 1 | 0 | 1.1817 | 0.1988 | 1 | 0 |
| $r = 0.7$ | ElasticNet | 1.0833 | 0.1112 | 0.7143 | 0 | 1.0947 | 0.0526 | 0.4102 | 0 |
| | BAR | 0.9500 | 0.0454 | 0.0614 | 0 | 0.9321 | 0.0094 | 0.0100 | 0 |
| | BALT | 0.9478 | 0.0486 | 0.0843 | 0 | 0.9244 | 0.0106 | 0.0151 | 0 |
| | Lasso | 1.1009 | 0.1581 | 0.5386 | 0 | 1.1922 | 0.1612 | 0.1783 | 0.3200 |
| | Ridge | 1.7034 | 0.6173 | 1 | 0 | 1.3892 | 0.2674 | 1 | 0 |
| $r = 0.9$ | ElasticNet | 1.0834 | 0.1955 | 0.7157 | 0 | 1.2049 | 0.2116 | 0.2768 | 0.2600 |
| | BAR | 0.9519 | 0.0783 | 0.0629 | 0 | 1.0622 | 0.1219 | 0.0113 | 0.5100 |
| | BALT | 0.9489 | 0.0858 | 0.0914 | 0 | 1.0519 | 0.1539 | 0.0472 | 0.5733 |

**Table 2**: Performance comparison for different methods under varying $r$ and $p$ for scenario 2.

| | | $p = 10$ | | | | $p = 50$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| $r$ | Method | MSE | MAE | FPR | FNR | MSE | MAE | FPR | FNR |
| | Lasso | 1.0725 | 0.0946 | 0.7375 | 0.0333 | 1.1202 | 0.0376 | 0.2659 | 0.1667 |
| | Ridge | 1.3141 | 0.2176 | 1 | 0 | 1.0026 | 0.1567 | 1.0000 | 0 |
| 0.5 | ElasticNet | 1.0357 | 0.0980 | 0.8125 | 0.0167 | 1.0938 | 0.0475 | 0.3830 | 0.0750 |
| | BAR | 0.9266 | 0.0984 | 0.3375 | 0.1167 | 0.8839 | 0.0232 | 0.0307 | 0.2667 |
| | ALIU | 0.9240 | 0.0966 | 0.3000 | 0.1167 | 0.9269 | 0.0229 | 0.0205 | 0.3250 |
| | Lasso | 1.0798 | 0.1276 | 0.7750 | 0.0917 | 1.1041 | 0.0498 | 0.3205 | 0.1417 |
| | Ridge | 1.4417 | 0.3160 | 1 | 0 | 1.1587 | 0.1953 | 1 | 0 |
| 0.7 | ElasticNet | 1.0443 | 0.1301 | 0.8125 | 0.0583 | 1.0722 | 0.0612 | 0.4341 | 0.1000 |
| | BAR | 0.9246 | 0.1213 | 0.4125 | 0.1750 | 0.9419 | 0.0220 | 0.0148 | 0.3583 |
| | BALT | 0.9438 | 0.1171 | 0.2625 | 0.2417 | 0.9387 | 0.0221 | 0.0136 | 0.3500 |
| | Lasso | 1.0801 | 0.1977 | 0.5875 | 0.1333 | 1.0867 | 0.0665 | 0.2739 | 0.2500 |
| | Ridge | 1.6723 | 0.6277 | 1 | 0 | 1.4368 | 0.2531 | 1 | 0 |
| 0.9 | ElasticNet | 1.0495 | 0.1911 | 0.7375 | 0.0833 | 1.1107 | 0.0917 | 0.4025 | 0.1750 |
| | BAR | 0.9349 | 0.1865 | 0.2125 | 0.2833 | 0.9645 | 0.0330 | 0.0519 | 0.4667 |
| | BALT | 0.9375 | 0.1827 | 0.1750 | 0.3167 | 0.9262 | 0.0364 | 0.0261 | 0.4500 |

**Table 3**: Performance comparison of methods under scenario 3.

| **Method** | **MSE** | **MAE** | **FPR** | **FNR** |
|---|---|---|---|---|
| Lasso | 27.0677 | 0.8634 | 0.5133 | 0.4300 |
| Ridge | 27.0772 | 0.2537 | 1.0000 | 0.0000 |
| ElasticNet | 27.1991 | 0.5788 | 0.5800 | 0.2267 |
| BAR | 24.3420 | 1.1254 | 0.1133 | 0.6267 |
| BALT | 24.5289 | 1.0680 | 0.1400 | 0.6700 |

**Table 4**: Performance comparison for different methods under varying $\rho$ values for scenario 4.

| | $\rho = 0.7$ | | | | $\rho = 0.9$ | | | |
|---|---|---|---|---|---|---|---|---|
| Method | MSE | MAE | FPR | FNR | MSE | MAE | FPR | FNR |
| Lasso | 10.7685 | 0.0483 | 0.0752 | 0.1680 | 10.0865 | 0.0574 | 0.0577 | 0.3480 |
| Ridge | 18.2374 | 0.0802 | 1 | 0 | 17.8263 | 0.0738 | 1 | 0 |
| ElasticNet | 10.7922 | 0.0478 | 0.1038 | 0.0600 | 10.3209 | 0.0466 | 0.0819 | 0.1000 |
| BAR | 8.2245 | 0.0668 | 0.0126 | 0.6520 | 8.9168 | 0.0721 | 0.0038 | 0.7600 |
| BALT | 7.9453 | 0.0625 | 0.0162 | 0.6080 | 8.4418 | 0.0712 | 0.0095 | 0.6960 |

**Table 5**: Performance comparison of methods under scenario 5.

| Method | MSE | MAE | FPR | FNR |
|---|---|---|---|---|
| Lasso | 1.2509 | 0.0347 | 0.1755 | 0.2500 |
| Ridge | 19.1155 | 0.1358 | 1 | 0 |
| ElasticNet | 1.6012 | 0.0484 | 0.2255 | 0.2333 |
| BAR | 0.9820 | 0.0135 | 0.0064 | 0.4500 |
| BALT | 0.8953 | 0.0166 | 0.0160 | 0.4333 |

# 5 Examples

In this section, we examine the predictive performance of five regularized regression estimators: Ridge, Lasso, Elastic Net (E-net), Broken Adaptive Ridge (BAR), and the proposed estimator, Broken Adaptive Liu-type (BALT). The evaluation is conducted using three benchmark datasets that have been widely employed in related studies on regularization methods.

The predictors were standardized such that

$$\frac{1}{n}\sum_{i=1}^{n} z_{ij}^2 = 1, \quad \text{for } j = 1, 2, \ldots, p,$$

and the response variable was centered. Each data set was partitioned into training and testing subsets, with 80% of the observations used for training and the remaining 20% reserved for testing. The fitting of the model and the selection of parameters were performed using a five-fold cross-validation on the training data. Specifically, the tuning parameters were selected from a discrete grid ranging from 0 to 50 for lambda, -10 to 10 for $d$ and the optimal value for each model was chosen based on the minimization of the average cross-validated mean squared error (MSE).

Once the optimal tuning parameter was identified, the corresponding model was refitted to the entire training set and its coefficients were used to predict the response in the test set. The generalization performance of each estimator was evaluated using the following metrics:

- **Mean Squared Prediction Error (MSPE)**:

$$\text{MSPE} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \left(y_i - \hat{y}_i\right)^2,$$

  where $y_i$ and $\hat{y}_i$ are the observed and predicted values, respectively, and $n_{\text{test}}$ is the number of test observations.
- **Mean Absolute Error (MAE)**:

$$\text{MAE} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \left|y_i - \hat{y}_i\right|.$$

In addition to prediction accuracy, we also evaluated model sparsity by recording the number of active variables (i.e., non-zero coefficients) selected by each estimator. This provides further insight into the ability of each method to perform variable selection and control the complexity of the model.

## 5.1 Example I: Prostate Cancer Data

The first data set used in this study involves clinical measurements related to prostate cancer. It is used to model the relationship between the level of prostate-specific antigen (PSA) and several explanatory variables obtained from male patients scheduled for radical prostatectomy [4, 13]. The objective is to understand how various clinical features influence PSA levels.

The regression model is formulated as follows:

$$y_i = \theta_0 + \theta_1 z_{i1} + \theta_2 z_{i2} + \cdots + \theta_8 z_{i8} + \epsilon_i, \quad i = 1, \ldots, 97, \tag{14}$$

where $y_i$ denotes the log-transformed PSA level for the $i$-th individual, and the predictors $x_{ij}$ are defined as:

- $z_{i1}$: Logarithm of cancer volume
- $z_{i2}$: Logarithm of prostate weight
- $z_{i3}$: Patient's age in years
- $z_{i4}$: Natural logarithm of benign prostatic hyperplasia volume
- $z_{i5}$: Indicator of seminal vesicle invasion
- $z_{i6}$: Logarithm of capsule penetration
- $z_{i7}$: Gleason score
- $z_{i8}$: Percentage of biopsy cores with a Gleason score of 4 or 5

**Table 6**: Comparison of coefficient estimates and performance metrics across models using prostate data

| **Predictors/COEF** | **Ridge** | **Lasso** | **Elastic Net** | **BAR** | **BALT** |
|---|---|---|---|---|---|
| lcavol | 0.4880 | 0.5147 | 0.4974 | 0.5601 | 0.5434 |
| lweight | 0.3163 | 0.1230 | 0.1121 | 0.4588 | 0.4587 |
| age | -0.0038 | 0.0179 | 0.0188 | | |
| lbph | 0.0821 | | | | |
| svi | 0.2674 | | | 0.6032 | 0.6102 |
| lcp | 0.0394 | | | | |
| gleason | 0.1131 | | | | |
| pgg45 | 0.0031 | 0.0069 | 0.0072 | | |
| MSPE | 0.7089 | 0.7792 | 0.7850 | 0.7077 | 0.7043 |
| MAE | 0.5518 | 0.6010 | 0.6036 | 0.5642 | 0.5640 |
| Active Set Size | 8 | 4 | 4 | 3 | 3 |
| Lambda | 19.2727 | 0.1597 | 0.1997 | 0.56 | 0.07 |
| $d$ | | | | | 35 |

Figure 2 displays the coefficient trajectories for four regularization methods **BAR**, **BALT**, **Lasso**, and **Elastic Net**—plotted as a function of the normalized $\ell_1$ norm, i.e. $||\boldsymbol{\beta}||_1 / \max ||\boldsymbol{\beta}||_1$. These paths provide valuable insights into the dynamics of variable selection and shrinkage behavior as the regularization strength is varied. BAR reveals a gradual evolution of the coefficient paths, indicating a relatively smooth transition in the inclusion of variables. Key predictors such as `lcavol`, `lweight`, and `svi` emerge early and maintain prominence throughout the path. The stability and interpretability of these trajectories affirm the robustness of BAR in capturing the essential structure of the data with moderate regularization. BALT exhibits more adaptive behavior, with sharper transitions and earlier activation of critical variables. The coefficient paths show a high degree of shrinkage for weaker predictors, while allowing dominant variables to reach larger magnitudes. This confirms BALT's flexibility and strength in selectively shrinking noise while preserving signal, enabled by its adaptive penalization scheme and the tuning parameter $d$. The abrupt transitions suggest strong prior adaptivity that emphasizes sparsity and parsimony.

The bottom-left panel shows that Lasso initiates variable inclusion at varying stages, with dominant predictors like `lcavol` entering early. However, some coefficients plateau prematurely, reflecting Lasso's tendency to overshrink when predictors are correlated. Additionally, Lasso fails to maintain stability in the presence of multicollinearity, often excluding informative variables in favor of those more weakly correlated with others. The bottom-right panel demonstrates that Elastic Net mitigates Lasso's limitations by allowing smoother transitions and incorporating more variables, thanks to its mixed $\ell_1$-$\ell_2$ penalty. However, this comes at the cost of slightly reduced sparsity and interpretability. The simultaneous activation of correlated predictors implies that Elastic Net is particularly useful in high-dimensional settings, albeit with less parsimony than BALT or BAR.

Overall, BALT and BAR offer more stable and interpretable coefficient paths, aligning with their favorable predictive performance shown earlier. BALT's sharper transitions and aggressive regularization make it highly suitable for sparse signal recovery. In contrast, Lasso and Elastic Net, while effective, may underperform in settings with complex correlation structures.

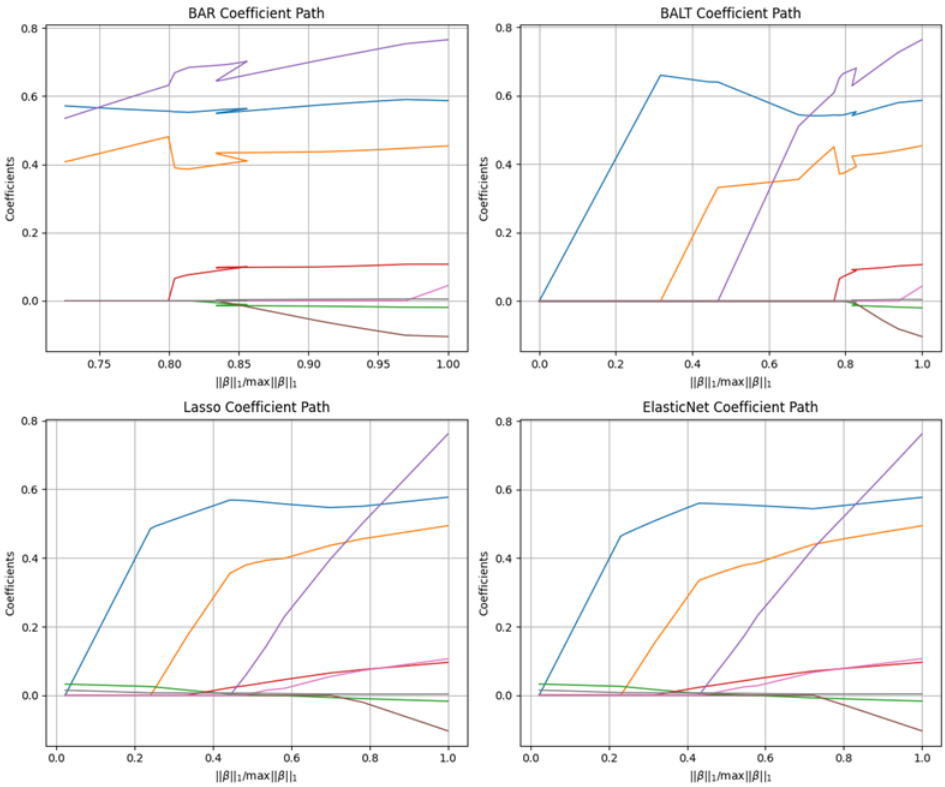


**Figure 2**: The coefficient paths of the BAR, BALT, lasso, and elastic net for the prostate data as a function of $\|\boldsymbol{\beta}\|_1 / \max \|\boldsymbol{\beta}\|_1$

Table 6 presents the coefficient estimates, predictive performance metrics, and model complexity indicators for five regression estimators—**Ridge**, **Lasso**, **Elastic Net**, **BAR**, and **BALT** applied to a prostate cancer dataset. This comparative analysis highlights the trade-offs between predictive accuracy, model sparsity, and interoperability. Among the evaluated methods, **BALT** achieves the lowest mean squared prediction error (MSPE = 0.7043), indicating superior generalization performance. This is closely followed by **BAR**, which also outperforms the conventional estimators such as Lasso and Elastic Net. The slightly higher MSPE values for Lasso (0.7792) and Elastic Net (0.7850) suggest that their regularization mechanisms may be suboptimal in this setting, potentially due to correlated predictors or weak signal strength. Both BALT and BAR select only three predictors—`lcavol`, `lweight`, and `svi`—highlighting their capacity for effective variable selection and highly interpretable offering models. This is particularly beneficial in clinical or policy applications where simplicity and clarity are essential. In contrast, Ridge regression retains all eight predictors due to the nature of the $\ell_2$ penalty, which, while preventing overfitting, results in less interpretable models. Lasso and Elastic Net strike a middle ground with four active predictors but fail to capture `svi`, a key variable retained by both BALT and BAR, which may partially explain their reduced performance.

The values of the regularization parameter ($\lambda$) further elucidate the behavior of each method. Ridge requires a high level of penalization ($\lambda = 19.27$), leading to uniform shrinkage. In contrast, BALT achieves optimal performance with a small $\lambda = 0.07$ and a large additional parameter $d = 35$, enabling more flexible and targeted regularization. This adaptivity is crucial in settings with noisy or high-dimensional features, where uniform penalization risks underfitting important variables or overfitting noise.

## 5.2 Example II: Electricity dataset

The Electricity dataset contains cost function data for 145 US electricity producers in 1955, with an additional 14 observations representing aggregate statistics [14]. For statistical analysis, only the first 145 observations should be used. The data set comprises eight variables:

- **cost**: Total production cost.
- **output**: Total output of electricity.
- **labor**: Wage rate of labor.
- **laborshare**: Cost share for labor.
- **capital**: Capital price index.
- **capitalshare**: Cost share for capital.
- **fuel**: Fuel price.
- **fuelshare**: Cost share for fuel.

**Table 7**: Comparison of Coefficients and Metrics Across Methods using Electricity Data

| Predictors/COEF | Ridge | Lasso | Elastic net | BAR | BALT |
|---|---|---|---|---|---|
| output | 17.5176 | 18.0407 | 18.7939 | 19.1808 | 19.1856 |
| labor | 1.4878 | 0.7546 | 1.0358 | 0.8476 | 0.8838 |
| laborshare | 0.6368 | 0.1056 | 0.8918 | 1.0718 | 1.1005 |
| capital | 0.5819 | . | 0.2986 | . | . |
| sat-co | -0.1989 | . | . | . | . |
| fuel | 1.3054 | 1.0306 | 1.4952 | 1.7531 | 1.7574 |
| fuelshare | -0.0543 | . | . | . | . |
| MSPE | 14.1204 | 14.142 | 14.0495 | 14.0504 | 14.0495 |
| MAE | 12.9897 | 13.0329 | 13.0521 | 13.0707 | 13.0695 |
| ACTIVE SET SIZE | 7 | 4 | 5 | 4 | 4 |
| $\lambda$ | 1.8789 | 0.7946 | 0.2964 | 26.4 | 22.4 |
| $d$ | . | . | . | . | 0 |

Table 7 shows that the Ridge regression produces smoothly shrunken coefficients, retaining all predictors in the model. In contrast, Lasso enforces sparsity by producing smaller coefficient estimates - evidenced by a lower estimate for *labor* - and by reducing the active set. Elastic Net strikes a balance between the behaviors of Ridge and Lasso. The performances of BAR and BALT are very similar; both methods select four predictors as determined by Lasso. Overall, MSPE and MAE are comparable across all methods, with BALT achieving the lowest MSPE. These results suggest that broken adaptive methods (BAR and BALT) can enhance variable selection without compromising predictive performance. Although traditional approaches like Ridge, Elastic Net, and Lasso remain effective, adaptive methods such as BAR and BALT provide a flexible alternative that better balances model complexity and accuracy.

## 5.3 Example III: Riboflavin Dataset

In this section, we examine a genomic dataset related to riboflavin (*vitamin B2*) production in *Bacillus subtilis*. This data set was initially investigated by Bühlmann [15] and is available through the `hdi` package in the `R` software environment. Further analyses have been conducted by Javanmard and Montanari [16], Zhang *et al.* [17], Genç and Özkale [6], among others.

The primary objective of these studies is to identify the genes that contribute to increased riboflavin production, enabling the creation of higher-yield bacterial strains. The data set consists of 71 observations and 4088 predictors, each corresponding to the logarithmic expression level of a specific gene. The response variable represents the logarithm of riboflavin production rates in *Bacillus subtilis*.

Table 8 summarizes the optimal tuning parameters, prediction errors, and model sparsity for five regression estimators (Ridge, Lasso, Elastic Net, BAR, and BALT) applied to the riboflavin dataset. Ridge regression yields the lowest prediction error (MSPE = 0.0566, MAD = 0.0429) but at the expense of including almost all available predictors (4088 active coefficients), which may compromise interpretability and increase computational burden in high-dimensional analyses. In contrast, Lasso and Elastic Net select substantially fewer predictors (52 and 53, respectively); however, they incur an elevated MSPE

**Table 8**: Performance metrics for various estimators on the riboflavin data.

| Estimators | MSPE | MAD | ACTIVE SETS | $\lambda$/d |
|---|---|---|---|---|
| Ridge | 0.0566 | 0.0429 | 4088 | 100 |
| Lasso | 0.1440 | 0.0429 | 52 | 0.0223 |
| Elastic-net | 0.1441 | 0.1082 | 53 | 0.0279 |
| BAR | 0.0688 | 0.0531 | 30 | 0.01 |
| BALT | 0.0568 | 0.0459 | 49 | 0.01/-2 |

(approximately 0.1440) and, in the case of Elastic Net, a higher MAD (0.1082), suggesting potential sensitivity to outliers or over-penalization.

Adaptive methods show considerable promise: the BAR estimator achieves competitive prediction performance (MSPE = 0.0688, MAD = 0.0531) using only 30 predictors. In comparison, the BALT estimator achieves MSPE (0.0568) and MAD (0.0459), similar to Ridge with a more parsimonious model (49 active predictors). The additional tuning parameter in BALT further enhances its flexibility in capturing the underlying sparsity.

These findings indicate that, while Ridge regression minimizes the prediction error, its dense model structure may not be suitable for applications requiring interpretability. Adaptive techniques, particularly BALT, offer a balanced alternative by maintaining a low prediction error with a markedly reduced active set, underscoring their potential in high-dimensional biological data analysis.

# 6 Concluding remarks

Ridge regression is a widely used technique for parameter estimation in linear models, offering improved stability over ordinary least squares, particularly in multicollinearity. However, Ridge regression cannot perform variable selection. Recent developments with the introduction of the Broken Adaptive Ridge (BAR) method have introduced mechanisms enabling shrinkage and variable selection within a Ridge-type framework. Similarly, the Liu-type estimator has been shown to offer competitive performance compared to Ridge regression by incorporating a biasing parameter, but fails to achieve sparsity. This study proposes the Broken Adaptive Liu-Type (BALT) estimator as a theoretically grounded and empirically validated solution to multicollinearity in multivariate regression models. By employing an adaptive, eigenstructure-driven shrinkage strategy, BALT effectively balances bias and variance while promoting sparsity through selective penalization. The proposed method extends the classical Liu-type estimators by introducing a broken adaptive framework, which allows for nuanced control over regularization intensity across orthogonal subspaces of the design matrix.

Our asymptotic analysis confirms that BALT satisfies both oracle and grouping properties under mild regularity conditions. These theoretical guarantees are supported by simulation results that demonstrate its superior performance in estimation precision, variable selection accuracy, and prediction error, relative to existing methods. The utility of BALT is further illustrated through applications to real datasets in medical diagnostics, economic modeling, and genomics, where it achieves competitive or improved predictive accuracy while maintaining model parsimony.

In summary, BALT contributes a flexible and efficient tool to the suite of penalized regression techniques, particularly suited for high-dimensional multicollinear settings. Future work may extend this framework to generalized linear models, structured sparsity regimes, and Bayesian implementations.

# Appendix: Proofs of the Main Results

To prove Theorems 1, we begin by introducing some convenient notation. Write

$$\theta \;=\; \begin{pmatrix} \alpha \\ \gamma \end{pmatrix},$$

where $\alpha \in \mathbb{R}^{q_n}$ and $\gamma \in \mathbb{R}^{\,p_n - q_n}$. Similarly, denote the $k$-th iterate by

$$\widehat{\theta}^{(k)} \;=\; \begin{pmatrix} \widehat{\alpha}^{(k)} \\ \widehat{\gamma}^{(k)} \end{pmatrix}.$$

For any candidate $\theta$, define the mapping

$$f(\theta) \;=\; \big(\mathbf{Z}^T\,\mathbf{Z} \;+\; \lambda_n D(\theta)\, I_n\big)^{-1}\big(\mathbf{Z}^T\,\mathbf{Z} \;-\; \lambda_n\, d\, F(\theta)\big)\,\theta^{\mathrm{LS}}. \tag{A.1}$$

We shall write

$$f(\tilde{\theta}) \;=\; \begin{pmatrix} \alpha^*(\tilde{\theta}) \\ \gamma^*(\tilde{\theta}) \end{pmatrix},$$

and for brevity set $\alpha^* = \alpha^*(\theta)$, $\gamma^* = \gamma^*(\theta)$ where no ambiguity arises.

Next, partition the inverse of the population Fisher information matrix, $\Sigma_n^{-1}$, as

$$\Sigma_n^{-1} \;=\; \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix},$$

with $B_{11} \in \mathbb{R}^{q_n \times q_n}$. Multiplying $(Z^TZ)^{-1}(Z^TZ + \lambda_n D(\theta) I_n)$ to (A.1) yields

$$\begin{pmatrix} \alpha^* - \alpha_0 \\ \gamma^* \end{pmatrix} \;+\; \frac{\lambda_n}{n}\begin{pmatrix} B_{11}D_1(\alpha)\,\alpha^* + B_{12}D_2(\gamma)\,\gamma^* \\ B_{21}D_1(\alpha)\,\alpha^* + B_{22}D_2(\gamma)\,\gamma^* \end{pmatrix} \;-\; \frac{\lambda_n\, d}{n\,\widetilde{\theta}_j}\begin{pmatrix} [B_{11},\, B_{12}]\,\theta^{\mathrm{LS}} \\ [B_{21},\, B_{22}]\,\theta^{\mathrm{LS}} \end{pmatrix} \;=\; (Z^TZ)^{-1}Z^T\varepsilon \tag{A.2}$$

where

$$D_1(\alpha) = \mathrm{diag}\big(\alpha_1^{-2}, \ldots, \alpha_{q_n}^{-2}\big), \qquad D_2(\gamma) = \mathrm{diag}\big(\gamma_1^{-2}, \ldots, \gamma_{p_n - q_n}^{-2}\big).$$

The following lemma captures the key bounds on $f$ over suitably small neighborhoods.

**Lemma 1.** Let $\{\delta_n\}$ be any sequence of positive reals satisfying $\delta_n \to \infty$ and $p_n\delta_n^2/\lambda_n \to 0$. Define

$$H_n = \big\{\theta \in \mathbb{R}^{p_n} : \|\theta - \theta_0\| \le \delta_n\sqrt{p_n/n}\big\}, \quad H_n^0 = \big\{\alpha \in \mathbb{R}^{q_n} : \|\alpha - \theta_{01}\| \le \delta_n\sqrt{p_n/n}\big\}.$$

Assume conditions (C1)–(C3) hold. Then with probability tending to 1,

1. $\displaystyle\sup_{\theta \in H_n} \frac{\|\gamma^*(\theta)\|}{\|\gamma\|} < \frac{1}{C_0}$ for some constant$C_0 > 1$;
2. $f\big(H_n\big) \subseteq H_n$.

*Proof of part (a).* Since $\lambda_n/\sqrt{n} \to 0$ and $p_n\delta_n^2/\lambda_n \to 0$, it follows that $\delta_n\sqrt{p_n/n} \to 0$. By (C2),

$$nE\big\|\,(Z^TZ)^{-1}Z^T\varepsilon\big\|^2 = \sigma^2 tr(\Sigma_n^{-1}) = O(p_n),$$

so $E\|\theta^{\mathrm{LS}} - \theta_0\| = O(p_n/n)$, and thus $\|\theta^{\mathrm{LS}} - \theta_0\| = O_p(\sqrt{p_n/n})$. Then from (A.2) we obtain

$$\sup_{\theta \in H_n}\left\|\gamma^* + \frac{\lambda_n}{n}\,B_{12}\,D_1(\alpha)\,\alpha^* \;+\; \frac{\lambda_n}{n}\,B_{22}\,D_2(\gamma)\,\gamma^* \;-\; \frac{\lambda_n\, d}{n\,\tilde{\theta}_j}\,[B_{21},\, B_{22}]\,\theta^{\mathrm{LS}}\right\| \;=\; O_p\big(\sqrt{p_n/n}\big). \tag{A.3}$$

Next, since $\|\alpha - \theta_{01}\| \le \delta_n\sqrt{p_n/n}$ and $\|\alpha^*\| \le \|\theta^{\mathrm{LS}}\| = O_p(b_{1n}\sqrt{p_n})$, assumptions (C2)–(C3) imply

$$\sup_{\theta\in H_n} \frac{\lambda_n}{n}\left\|B_{12}\, D_1(\alpha)\,\alpha^*\right\| \;\le\; \frac{\lambda_n}{n}\|B_{12}\| \sup_{\theta\in H_n}\left\|D_1(\alpha)\,\alpha^*\right\| \;\le\; \sqrt{2}C\frac{\lambda_n}{n}\frac{b_{1n}}{b_{0n}^2}\sup_{\theta\in H_n}\|\alpha^*\| \;=\; O_p\big(\sqrt{p_n/n}\big), \tag{A.4}$$

where we used $\|B_{12}\| \le \sqrt{2}C$ (since $\|B_{12}B_{21}\| - \|B_{11}^2\| \le \|B_{11}^2 + B_{12}B_{21}\|^2 \le \|\Sigma_n^{-1}\| \le C^2$). Combining (A.3) and (A.4) yields

$$\sup_{\theta\in H_n}\left\|\gamma^* + \frac{\lambda_n}{n\,\tilde\theta_j^2}\,B_{22}\,D_2(\gamma)\,\gamma^* - \frac{\lambda_n\, d}{n\,\tilde\theta_j}\,[B_{21},\, B_{22}]\,\theta^{\mathrm{LS}}\right\| \;=\; O_p\big(\sqrt{p_n/n}\big) \tag{A.5}$$

Note that $B_{22}$ is positive definite, and by the singular value decomposition, we can write

$$B_{22} = \sum_{i=1}^{p_n-q_n} \tau_{2i}\mathbf{u}_{2i}\mathbf{u}_{2i}^T,$$

where $\tau_{2i}$ and $\mathbf{u}_{2i}$ are eigenvalues and eigenvectors of $B_{22}$. Then, since $1/C < \tau_{2i} < C$ for some constant $C > 1$, we have

$$\begin{aligned}
\frac{\lambda_n}{n}\|B_{22}D_2(\gamma)\,\gamma^*\| &= \frac{\lambda_n}{n}\left\|\sum_{i=1}^{p_n-q_n}\tau_{2i}\mathbf{u}_{2i}\mathbf{u}_{2i}^T D_2(\gamma)\,\gamma^*\right\| \\
&\ge \frac{\lambda_n}{n}\left(\sum_{i=1}^{p_n-q_n}\tau_{2i}^2\|\mathbf{u}_{2i}^T D_2(\gamma)\,\gamma^*\|^2\right)^{1/2} \\
&\ge \frac{1}{C}\left(\frac{\lambda_n}{n}\right)\left(\sum_{i=1}^{p_n-q_n}\|\mathbf{u}_{2i}^T D_2(\gamma)\,\gamma^*\|^2\right)^{1/2} \\
&= \frac{1}{C}\frac{\lambda_n}{n}\|D_2(\gamma)\,\gamma^*\|.
\end{aligned}$$

This, together with (A.5) and condition (C2) for BALT, implies that with probability tending to 1,

$$\frac{1}{C}\frac{\lambda_n}{n}\|D_2(\gamma)\,\gamma^*\| - \|\gamma^*\| \le \delta_n\sqrt{p_n/n}. \tag{A.6}$$

Define

$$d_{\gamma^*/\gamma} = \left(\frac{\gamma_1^*}{\gamma_1}, \ldots, \frac{\gamma_{p_n-q_n}^*}{\gamma_{p_n-q_n}}\right)^T.$$

Because $\|\gamma\| \le \delta_n\sqrt{p_n/n}$ within $H_n$, we have

$$\begin{aligned}
\frac{1}{C}\frac{\lambda_n}{n}\|D_2(\gamma)\,\gamma^*\| &= \frac{1}{C}\frac{\lambda_n}{n}\left(\|D_2(\gamma)^{1/2}d_{\gamma^*/\gamma}\|\right) \\
&\ge \frac{1}{C}\frac{\lambda_n}{n}\frac{\sqrt{n}}{\delta_n\sqrt{p_n}}\|d_{\gamma^*/\gamma}\|.
\end{aligned} \tag{A.7}$$

Meanwhile, we observe that

$$\|\gamma^*\| = \|D_2(\gamma)^{-1/2}d_{\gamma^*/\gamma}\| \le \frac{\delta_n\sqrt{p_n}}{\sqrt{n}}\|d_{\gamma^*/\gamma}\|. \tag{A.8}$$

Combining (A.6)–(A.8), we obtain that with probability tending to 1,

$$\|d_{\gamma^*/\gamma}\| \leq \frac{1}{\lambda_n/(p_n\delta_n^2 C) - 1} < \frac{1}{C_0} \quad \text{for some constant } C_0 > 1,$$

provided that $\lambda_n/(p_n\delta_n^2) \to \infty$.

Furthermore, with probability tending to 1,

$$\|\gamma^*\| \leq \|d_{\gamma^*/\gamma}\| \cdot \max_{1 \leq j \leq p_n - q_n} |\gamma_j| \leq \|\gamma\| \|d_{\gamma^*/\gamma}\| \leq \frac{1}{C_0}\|\gamma\|.$$

Thus, we have established that $\|\gamma^*\|$ is strictly smaller than $\|\gamma\|$ with high probability, completing the proof of part (a) for BALT.

□

We now turn to establishing part (b). First, observe from (A.8) and (A.9) that, as $n \to \infty$,

$$\Pr\left(\|\theta^*\| \leq \delta_n\sqrt{p_n/n}\right) \to 1. \tag{A.10}$$

Using (A.2), it follows that

$$\sup_{\theta \in \mathcal{H}_n} \left\| \alpha^* - \theta_{01} + \lambda_n B_{11} D_1(\alpha)\alpha^*/n) + \lambda_n B_{12} D_2(\gamma)\gamma^*/n - \frac{\lambda_n d}{n\,\tilde{\theta}_j} \left[B_{11},\, B_{12}\right] \theta^{\mathrm{LS}} \right\| = O_p\left(\sqrt{p_n/n}\right). \tag{A.11}$$

Following similar reasoning as in (A.4), we find that

$$\sup_{\theta \in \mathcal{H}_n} \|\lambda_n B_{11} D_1(\alpha)(\alpha^*)/n)\| = o_p\left(\sqrt{p_n/n}\right). \tag{A.12}$$

Additionally, with high probability,

$$\sup_{\theta \in \mathcal{H}_n} \|\lambda_n B_{12} D_2(\gamma)\gamma^*/n\| \leq \frac{\lambda_n}{n} \sup_{\theta \in \mathcal{H}_n} \|D_2(\gamma)\| \cdot \|B_{12}\| \leq 2\sqrt{2}C^2\delta_n\sqrt{p_n/n}, \tag{A.13}$$

where the final inequality makes use of (A.6), (A.10), and the bound $\|B_{12}\| \leq \sqrt{2}C$.

Therefore, combining (A.11)–(A.13), we have with high probability:

$$\|\theta^* - \theta_{01}\| \leq \left(2\sqrt{2}C^2 + 1\right)\delta_n n^{-1/2}\sqrt{p_n}. \tag{A.14}$$

Since $\delta_n\sqrt{p_n}/\sqrt{n} \to 0$ as $n \to \infty$, it follows that

$$\Pr(\theta^* \in \mathcal{H}_{n1}) \to 1. \tag{A.15}$$

Putting together results (A.10) and (A.15) completes the justification for part (b). □

**Lemma 2.** Assume conditions (C1)–(C3) hold. For any $q_n$-dimensional vector $\boldsymbol{a}_n$ with $\|\boldsymbol{a}_n\| \leq 1$, define $s_n^2 = \sigma^2 \boldsymbol{a}_n^T \boldsymbol{\Sigma}_{n1} \boldsymbol{a}_n$ as in Theorem 1. Let

$$f(\boldsymbol{\alpha}) = \left(\mathbf{Z}_1^T\mathbf{Z}_1 + \lambda_n D_1(\alpha))\right)^{-1} \left(\mathbf{Z}_1^\top\mathbf{Z}_1 - \lambda_n d\,\mathrm{diag}(\tilde{\theta}_1^{-1}, \ldots, \tilde{\theta}_{q_n}^{-1})\right) (\boldsymbol{Z}_1^T\boldsymbol{Z}_1)^{-1}\boldsymbol{Z}_1^T\boldsymbol{y}, \tag{A.16}$$

Then, as $n \to \infty$:

(a) The function $f$ is a contraction on the set $\mathcal{B}_n = \left\{\boldsymbol{\alpha} \in \mathbb{R}^{q_n} : \|\boldsymbol{\alpha} - \boldsymbol{\theta}_0\| \leq \delta_n,\ \delta_n = o(\sqrt{p_n/n})\right\}$.

(b) Let $\hat{\boldsymbol{\alpha}}^\circ = f(\hat{\boldsymbol{\alpha}})$ denote the unique fixed point of $f$. Then,

$$\frac{\sqrt{n}}{s_n}\boldsymbol{a}_n^T(\hat{\boldsymbol{\alpha}}^\circ - \boldsymbol{\theta}_0) \to \mathcal{N}(0,1).$$

**Proof.** Rewrite $f(\boldsymbol{\alpha})$ as:

$$f(\alpha) - \theta_{01} + \lambda_n \Sigma_{n1}^{-1} D_1(\alpha) f(\alpha)/n - \frac{\lambda_n d}{n\tilde{\theta}} \Sigma_{n1}^{-1} \theta^{LS} = (Z_1^T Z_1)^{-1} Z_1^T \varepsilon.$$

Thus,

$$\sup_{\alpha \in \mathcal{B}_n} \left\| f(\alpha) - \theta_{01} + \lambda_n \Sigma_{n1}^{-1} D_1(\alpha) f(\alpha)/n - \frac{\lambda_n d}{n\tilde{\theta}} \Sigma_{n1}^{-1} \theta^{LS} \right\| = \mathcal{O}_p(\sqrt{q_n/n}). \tag{A.17}$$

Similar to (A.4), it can be shown that

$$\sup_{\alpha \in \mathcal{B}_n} \left\| \lambda_n \Sigma_{n1}^{-1} D_1(\alpha) f(\alpha)/n - \frac{\lambda_n d}{n\tilde{\theta}} \Sigma_{n1}^{-1} \theta^{LS} \right\| = O_p(\sqrt{q_n/n}). \tag{A.18}$$

It follows from (A.17) and (A.18) that

$$\sup_{\alpha \in \mathcal{B}_n} \|f(\alpha) - \theta_{01}\| \le \delta_n \sqrt{q_n/n}, \tag{A.19}$$

where $\delta_n$ is a sequence of real numbers satisfying $\delta_n \to \infty$ and $\delta_n \sqrt{q_n/n} \to 0$. This implies that, as $n \to \infty$,

$$\Pr(f(\alpha) \in \mathcal{B}_n) \to 1.$$

In other words, $f$ is a mapping from the region $\mathcal{B}_n$ to itself.

Rewrite (A.16) as $(Z_1^T Z_1 + \lambda_n D_1(\alpha)) f(\alpha) = \left( \mathbf{Z}_1^\top \mathbf{Z}_1 - \lambda_n d \operatorname{diag}(\tilde{\theta}_1^{-1}, \ldots, \tilde{\theta}_{q_n}^{-1}) \right) (\boldsymbol{Z}_1^T \boldsymbol{Z}_1)^{-1} X_1^T y$ and then differentiate it with respect to $\alpha$, we have

$$\{\Sigma_{n1} + \lambda_n D_1(\alpha)/n\} \frac{\partial f(\alpha)}{\partial \alpha^T} + (\lambda_n/n) \times \operatorname{diag}\left\{-2f(\alpha)/\alpha_j^2\right\} = 0,$$

where $f(\alpha) = \partial f(\alpha)/\partial \alpha^T$ and $\operatorname{diag}\left\{-2f(\alpha)/\alpha_j^2\right\} = \operatorname{diag}\left(-2f_1(\alpha)/\alpha_1^2, \ldots, -2f_{q_n}(\alpha)/\alpha_{q_n}^2\right)$. This, together with the assumption $\lambda_n/\sqrt{n} = o_n(1)$, implies that

$$\sup_{\alpha \in \mathcal{B}_n} \|\Sigma_{n1} + \lambda_n D_1(\alpha)/n\| \cdot \left\| \frac{\partial f(\alpha)}{\partial \alpha^T} \right\| \le \frac{2\lambda_n}{n} \sup_{\alpha \in \mathcal{B}_n} \left\| \operatorname{diag}\left\{ f_j(\alpha)/\alpha_j^2 \right\} \right\| = o_p(1). \tag{A.20}$$

Note that $\Sigma_{n1}$ is positive definite. Write $\Sigma_{n1} = \sum_{i=1}^{q_n} \tau_{1i} \mathbf{u}_{1i} \mathbf{u}_{1i}^T$, where $\tau_{1i}$ and $\mathbf{u}_{1i}$ are eigenvalues and eigenvectors of $\Sigma_{n1}$. Then, by (C2), $\tau_i \in (1/C, C)$ for all $i$ and

$$\begin{aligned} \|\Sigma_{n1} f(\alpha)\| &= \sup_{\|\mathbf{x}\|=1} \|\Sigma_{n1} f(\alpha)\mathbf{x}\| = \sup_{\|\mathbf{x}\|=1} \left\| \sum_{i=1}^{q_n} \lambda_{1i} \mathbf{u}_{1i} \mathbf{u}_{1i}^T f(\alpha)\mathbf{x} \right\| \\ &\le \sup_{\|\mathbf{x}\|=1} \left( \sum_{i=1}^{q_n} \lambda_{1i}^2 \left\| \mathbf{u}_{1i}^T f(\alpha)\mathbf{x} \right\|^2 \right)^{1/2} \le \sup_{\|\mathbf{x}\|=1} \frac{1}{C} \left( \sum_{i=1}^{q_n} \left\| \mathbf{u}_{1i}^T f(\alpha)\mathbf{x} \right\|^2 \right)^{1/2} \\ &= \sup_{\|\mathbf{x}\|=1} \frac{1}{C} \|f(\alpha)\mathbf{x}\| = \frac{1}{C} \|f(\alpha)\|. \end{aligned} \tag{A.21}$$

Therefore, it follows from $\alpha \in \mathcal{B}_n$, (A.21), and (C2) that

$$\|\Sigma_{n1} + \lambda_n D_1(\alpha)/n\| \cdot \left\| \frac{\partial f(\alpha)}{\partial \alpha^T} \right\| \ge \|\Sigma_{n1} f(\alpha)\| - \|\lambda_n D_1(\alpha) f(\alpha)/n\| \ge \frac{1}{C} \|f(\alpha)\| - \frac{\lambda_n}{n} \alpha_{\min}^{-2} \|f(\alpha)\|.$$

This, together with (A.20) and (C2), implies that

$$\sup_{\alpha\in\mathcal{B}_n} \|f(\alpha)\| = o_p(1). \tag{A.22}$$

Finally, the conclusion in part (a) follows from (A.19) and (A.22).

To demonstrate part (b), consider the following decomposition:

$$\begin{aligned} n^{1/2}\mathbf{S}_n^{-1}\mathbf{a}_n(\hat{\boldsymbol{\alpha}}^\circ - \boldsymbol{\theta}_0) = n^{1/2}\mathbf{S}_n^{-1}\mathbf{a}_n^\top \left[(\boldsymbol{\Sigma}_{n1} + \lambda_n \mathbf{D}_1(\hat{\boldsymbol{\alpha}}^\circ)/n)^{-1} M - \mathbf{I}_{q_n}\right] \boldsymbol{\theta}_{01} \\ + n^{1/2}\mathbf{S}_n^{-1}\mathbf{a}_n^\top \left(\boldsymbol{\Sigma}_{n1} + \lambda_n \mathbf{D}(\hat{\boldsymbol{\alpha}}^\circ)/n\right)^{-1} M\Sigma_{n1}\mathbf{X}_1^\top \boldsymbol{\varepsilon} \equiv I_1 + I_2. \end{aligned} \tag{A.23}$$

Using the first-order resolvent expansion, namely

$$(\mathbf{H}+\Delta)^{-1} = \mathbf{H}^{-1} - \mathbf{H}^{-1}\Delta(\mathbf{H}+\Delta)^{-1},$$

we rewrite the first term $I_1$ as:

$$I_1 = -\mathbf{S}_n^{-1}\mathbf{a}_n^\top \boldsymbol{\Sigma}_n^{-1} \frac{1}{\sqrt{n}} \mathbf{D}_1(\hat{\boldsymbol{\alpha}}^\circ) \left(\boldsymbol{\Sigma}_n + \lambda_n \mathbf{D}_1(\hat{\boldsymbol{\alpha}}^\circ)/n)\right)^{-1} M\boldsymbol{\Sigma}_{n1}\boldsymbol{\theta}_{01}.$$

Assuming conditions (C2) and (C3), we can bound the norm of $I_1$ as:

$$\|I_1\| \le \frac{\lambda_n}{\sqrt{n}} s_n^{-1} b_{0n}^{-2} \left\|\boldsymbol{\Sigma}_n^{-1} M\boldsymbol{\theta}_0\right\| = O_p\left(\frac{\lambda_n b_{1n}}{b_{0n}^2}\sqrt{\frac{q_n}{n}}\right) \to 0. \tag{A.24}$$

Next, express $\mathbf{Z}_1^\top = (\tilde{\mathbf{w}}_1, \ldots, \tilde{\mathbf{w}}_n)$, and again use the resolvent expansion to show that:

$$I_2 = \frac{\mathbf{S}_n^{-1}}{\sqrt{n}} \sum_{i=1}^n \mathbf{a}_n^\top M\boldsymbol{\Sigma}_n^{-1}\tilde{\mathbf{w}}_i\varepsilon_i + O_p(1), \tag{A.25}$$

which converges in distribution to a normal distribution $\mathcal{N}(0,1)$, as stated by the Lindeberg–Feller Central Limit Theorem.

Combining equations (A.23), (A.24), and (A.25), we complete the proof of part (b).

*Proof of Theorem 1.* Note that for the initial Liu-type estimator estimator $\widehat{\theta}^{(0)}$ defined by $\widehat{\theta}^{(0)} = (Z^TZ + \lambda_n I_n)^{-1}(Z^TZ - \lambda_n d)\,\theta^{\mathrm{LS}}$, we have

$$\widehat{\theta}^{(0)} - \theta_0 = \left\{(\Sigma_n + \xi_n I_{p_n}/n)^{-1}M - I_{p_n}\right\}\beta_0 \;+\; (\Sigma_n + \xi_n I_{p_n}/n)^{-1}M\Sigma_n^{-1}\frac{1}{n}X^T\varepsilon \;\equiv\; T_1 + T_2.$$

where $M = (\Sigma_n - \xi_n d I_{p_n}/n)$ By the first-order resolvent expansion, $\xi_n/\sqrt{n} \to 0$, and $d_n = O(1)$,

$$\|T_1\| = \left\|-\xi_n \tfrac{d_n}{n}(\Sigma_n + \tfrac{\xi_n}{n} I)\,\beta_0\right\| \le C\frac{\xi_n d_n}{n}\|\beta_0\| = O_p(\sqrt{p_n/n}), \quad \|T_2\| = O_p(\sqrt{p_n/n}),$$

so that $\|\widehat{\theta}^{(0)} - \theta_0\| = O_p(\sqrt{p_n/n})$.

This, combined with part (a) of Lemma 1, implies that

$$\Pr\Big(\lim_{k\to\infty} \widehat{\gamma}^{(k)} = 0\Big) \;\to\; 1. \tag{A.26}$$

Hence, to prove part (i) of Theorem 1, it suffices to show that

$$\Pr\Big(\lim_{k\to\infty} \|\widehat{\alpha}^{(k)} - \alpha^*\| = 0\Big) \;\to\; 1, \tag{A.27}$$

where $\alpha^*$ is the fixed point of $f(\alpha)$ defined in Lemma 2(b). Define $\gamma^* = 0$ if $\gamma = 0$. It is easy to see from (A.2) that for any $\alpha \in B_n$

$$\lim_{\gamma \to 0} \gamma^*(\alpha, \gamma) = 0. \tag{A.28}$$

For any $\alpha \in B_n$,

$$\lim_{\gamma \to 0} \alpha^*(\alpha, \gamma) = \left\{ Z_1^T Z_1 + \lambda_n D_1(\alpha) \right\}^{-1} M Z_1^T y = f(\alpha). \tag{A.29}$$

Therefore, $g$ is continuous and thus uniformly continuous on the compact set $\theta \in H_n$. This, together with (A.26) and (A.29), implies that, as $k \to \infty$,

$$\eta_k \equiv \sup_{\alpha \in B_n} \left\| f(\alpha) - \alpha^*(\alpha, \widehat{\gamma}^{(k)}) \right\| \ \to \ 0 \tag{A.30}$$

with probability tending to 1. Note that

$$\|\widehat{\alpha}^{(k+1)} - \alpha^*\| = \left\|\alpha^*(\widehat{\theta}^{(k)}) - \alpha^*\right\| \le \left\|\alpha^*(\widehat{\theta}^{(k)}) - f(\widehat{\alpha}^{(k)})\right\| + \left\|f(\widehat{\alpha}^{(k)}) - \alpha^*\right\| \le \eta_k + \tfrac{1}{C}\|\widehat{\alpha}^{(k)} - \alpha^*\|.$$

Let $a_k = \|\widehat{\alpha}^{(k)} - \alpha^*\|$ for every integer $k \ge 0$. From (A.30) we can inductively show that with probability tending to 1, for any $\epsilon > 0$ there exists $N$ such that for every integer $k > N$,

$$a_{k+1} \ \le \ \frac{a_1 + \eta_1 + \cdots + \eta_N}{C^{\,k-N}} \ + \ \frac{1 - (1/C)^{\,k-N}}{1 - 1/C}\,\epsilon,$$

and the right-hand term tends to 0 as $k \to \infty$. This proves (A.27).

Therefore, it follows immediately from (A.26) and (A.27) that with probability tending to 1,

$$\lim_{k \to \infty} \theta^{(k)} = \lim_{k \to \infty} (\widehat{\alpha}^{(k)T}, \widehat{\gamma}^{(k)T})^T = (\alpha^{*T}, 0^T)^T,$$

which completes the proof of part (i). This, in addition to part (b) of Lemma 2, proves part (ii) of Theorem 1. □

# References


[1] Akaike H, A new look at the statistical model identification, IEEE Trans. Automat. Contr 19 716–723, 1974.

[2] Arashi M, Asar Y, Yüzbaşı B, SLASSO: A scaled LASSO for multicollinear situations, *Journal of Statistical Computation and Simulation*, 91(15):3170–3183, 2021. Taylor & Francis.

[3] Chen J, Chen Z, Extended bayesian information criteria for model selection with large model

[4] Dai L, Chen K, Sun Z, Liu Z, Li G. Broken adaptive ridge regression and its asymptotic properties. *J Multivar Anal*. 2018 Nov;168:334-351, 2018.

[5] Fan J, Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*. 96, 456:1348-1360, 2001.

[6] Genç M, Özkale MR, Usage of the GO estimator in high dimensional linear models, *Computational Statistics*, 36(1):217–239, 2021. Springer.

[7] Knight K, Fu W, Asymptotics for lasso-type estimators, *Annals of Statistics*, pp. 1356–1378, 2000. JSTOR.

[8] Lukman AF, Ayinde K, Binuomote S, Clement OA, Modified ridge-type estimator to combat multicollinearity: Application to chemical data, *Journal of Chemometrics*, 33(5):e3125, 2019. Wiley Online Library.

[9] Lukman AF, Allohibi J, Jegede SL, Adewuyi ET, Oke S, Alharbi AA, Kibria–Lukman-Type Estimator for Regularization and Variable Selection with Application to Cancer Data, *Mathematics*, 11(23):4795, 2023. MDPI.

[10] Genç M, Lukman A, Weighted LAD-Liu-LASSO for robust estimation and sparsity, *Computational Statistics*, pp. 1–30, 2025. Springer.

[11] Genç M, A new double-regularized regression using Liu and lasso regularization, *Computational Statistics*, 37(1):159–227, 2022. Springer.

[12] Yüzbaşı B, Arashi M, Ahmed SE, Shrinkage estimation strategies in generalised ridge regression models: low/high-dimension regime, *Int. Stat. Rev.*, 88(1):229–251, 2020.

[13] Zou H, Hastie T, Regularization and variable selection via the elastic net, *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(2):301–320, 2005.

[14] Greene WH, *Econometric Analysis*, Pretence Hall, 2003.

[15] Bühlmann P, Kalisch M, Meier L, High-dimensional statistics with a view toward applications in biology, Annu. Rev. Stat. Appl., 1(1):255–278, 2014.

[16] Javanmard A, Montanari A, Confidence intervals and hypothesis testing for high-dimensional regression, J. Mach. Learn. Res., 15(1):2869–2909, 2014.

[17] Zhang C, Wu Y, Zhu M, Pruning variable selection ensembles, Stat. Anal. Data Min.: The ASA Data Sci. J., 12(3):168–184, 2019.

[18] Goeman, J. J., Meijer, R. J., & Chaturvedi, N. Penalized estimation methods for zero-inflated regression models. *Statistical Modelling*, 14(3), 215-237, 2014.

[19] Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55-67, 1970.

[20] Kibria, B. M. G. Performance of some new ridge regression estimators. *Communications in Statistics - Simulation and Computation*, 32(2), 419–435, 2003.

[21] Lee, A. H., & Silvapulle, M. J. Ridge estimation in logistic regression. *Communications in Statistics-Simulation and Computation*, 17(4), 1231-1257, 1988.

[22] Le Cessie, S., & Van Houwelingen, J. C. Ridge estimators in logistic regression. *Applied Statistics*, 41(1), 191-201, 1992.

[23] Liu, K. A new class of biased estimate in linear regression. *Communications in Statistics - Theory and Methods*, 22(2), 393–402, 1993.

[24] Liu, K. Using Liu-type estimator to combat collinearity. *Communications in Statistics—Theory and Methods*, 32(5), 1009–1020, 2003.

[25] Lukman, A. F., Adewuyi, E., Månsson, K., & Kibria, B. M. G. A new estimator for the multicollinear Poisson regression model: simulation and application. *Scientific Reports*, 11(1), 3732, 2021.

[26] Lukman, A. F., Aladeitan, B., Ayinde, K., & Abonazel, M. R. Modified ridge-type for the Poisson regression model: simulation and application. *Journal of Applied Statistics*, 49(8), 2124–2136, 2022.

[27] Massy, W. F. Principal Components Regression in Exploratory Statistical Research. *Journal of the American Statistical Association*, 60(309), 234–256, 1965.

[28] Månsson, K., & Shukur, G. A Poisson ridge regression estimator. *Economic Modelling*, 28(4), 1475-1481, 2011.

[29] Månsson, K. On ridge estimators for the negative binomial regression model. *Economic Modelling*, 29(2), 178-184, 2012.

[30] Montgomery, D. C., Peck, E. A., & Vining, G. G. Introduction to Linear Regression Analysis. *Wiley*, 2012.

[31] Özkale, M. R., & Kaciranlar, S. The restricted and unrestricted two-parameter estimators. *Communications in Statistics—Theory and Methods*, 36(15), 2707–2725, 2007.

[32] Zou H, Hastie T, Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 2005. Oxford University Press.

[33] Zou H, The adaptive lasso and its oracle properties, *Journal of the American Statistical Association*, 101(476):1418–1429, 2006. Taylor & Francis.