; revised .

Measurement Error In High-Dimensional Data

Herman F. Tesso Tassang *, **Prof. Geoges Nguefack-Tsague** †

[1] Department of Mathematics, University of Yaounde 1, BP:812 Yaounde-Cameroon
[2] Department of Public Health, Faculty of Medicine and Biomedical Sciences, UY1, PB 8550 Yaounde, Cameroon

Herman (e-mail: herman@aims.ac.za).

ABSTRACT In many important statistical applications, there are a large number of variables (covariates) p compared to the number of observations (sample size) n . These types of data are called high-dimensional data. Previous studies have shown that common statistical methods of analysis are not suitable for this type of data. While much work has been done on high-dimensional regression with clean data, it is important to consider corrupted data in real-world applications, such as genomics, where measurement error cannot be ignored. Therefore, it is necessary to highlight appropriate statistical methods for handling high-dimensional data, including cases where covariates are mismeasured. The purpose of this study is to introduce regularization methods (penalized linear regression) for reducing high-dimensional data and their variants that can account for measurement errors in covariates. In this paper, we evaluate four penalization methods: ridge regression, Lasso, Dantzig selector, and the Elastic net for model fitting. We also present their respective variants that accommodate measurement errors and compare them in a particular context of additive error with no correlation among measurement errors. Our evaluation focuses on practical applications and includes both simulated and real examples of high-dimensional datasets.

INDEX TERMS High-dimensional data; measurement error; penalized regression; Lasso; Dantzig Selector; Elastic net; ridge regression; Convex conditional Lasso; Non-convex lasso; Matrix Uncertainty Selector.

I. INTRODUCTION

THIS paper is about measurement error in high-dimensional data. In recent decades, technological progress has led to a great abundance of data in many scientific fields. For example, in genetics, a new framework has been developed in which the number of variables **p** is larger than the number of observations **n** (high-dimensional data). High-dimensional data analysis has experienced tremendous growth in popularity, and a plethora of methods have been proposed for statistical modeling and inference in high-dimensional data. Penalized regression methods such as ridge regression [15], Lasso [24], and Dantzig selector [3] are particularly effective in this context.

In almost all disciplines, it may not be possible to observe a variable accurately for some reason. Therefore, it is necessary to work with an error-prone version of that variable. Any measurement process can be affected by errors, usually due to the measuring instrument or the sampling process. The consequences of ignoring measurement error, many of which have been known for some time, can range from nonexistent to rather dramatic. Throughout this work, attention is given to the effects of measurement error on analyses that ignore it. This is mainly because the majority of researchers do not account for measurement error, even if they are aware of its presence and potential impact. In part, this is because the information or extra data needed to correct for measurement error may not be available. Typically, when measurement error creeps into the data, there are three main reasons why measurement error cannot be ignored: it can cause bias in parameter estimation [2], interfere with variable selection [23], and lead to a loss of power [4], resulting in trouble in detecting relationships among variables. Results on the bias of naive estimators often provide the added bonus of suggesting a correction method.

Applying high-dimensional regression methods that do not correct for measurement errors results in faulty inference, as demonstrated for the Lasso [21]. Consequently, correction for measurement error in penalized regression has recently been studied by various authors. Examples include "Ridge regression approach to measurement error" [21], Non-Convex Lasso (NCL) by Loh and Wainwright [23], the Convex Conditional Lasso (CoCoLasso) of Datta and Zou [7], and the Matrix Uncertainty Selector proposed by Rosenbaum and Tsybakov (MUS) [19].

The organization of this paper is as follows: **Section.** II and III present high-dimensional data together with potential challenges when analyzing the latter, along with some statistical methods one may use to handle this kind of dataset. **Section.**V introduces measurement error in regression theory, provides an overview of the consequences of measurement error in linear regression, and introduces some correction methods. **Section.**VI describes the behavior of measurement error in high-dimensional regression and introduces some high-dimensional approaches (methods) to correct for measurement error in the high-dimensional context. Both real and simulated data are used for illustrations.

II. Introduction to High-Dimensional Data

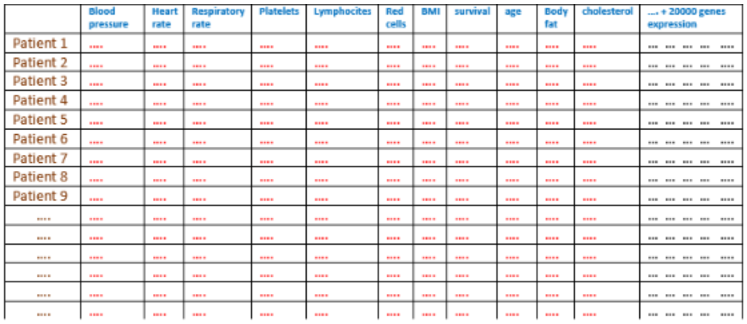
High-dimensional data is defined as data in which the number of features (*variables observed*), **p**, is close to or larger than the number of observations (or *data points*), **n**. The opposite is **low-dimensional data**, in which the number of observations, **n**, far outnumbers the number of features, **p**.

A related concept is **wide data**, which refers to data with numerous features irrespective of the number of observations; similarly, **tall data** is often used to denote data with a large number of observations. This concept should not be confused with the notion of **big data**, which is data that contains greater *variety*, arrives in increasing *volumes*, and has more *velocity*, known as the three **Vs** (visit, https://www.oracle.com/big-data/what-is-big-data/).

High-dimensional datasets are becoming more common in many scientific fields as new automated data collection techniques have been developed. An example in the biological sciences may include *data collected from hospital patients recording symptoms, blood test results, behaviors, and general health*, resulting in datasets with a large number of features.

An example of what high-dimensional data might look like in a biomedical study is shown in figure 1 below. Here are examples of descriptions of research questions whose associated datasets can be considered as high-dimensional data:

- Predicting patient blood pressure using: *cholesterol level in blood, age, and BMI, as well as information on 200,000 single nucleotide polymorphisms from 100 patients.*
- Predicting the probability of a patient's cancer progressing using: *gene expression data from 20,000 genes, as well as data associated with general patient health (age, weight, BMI, blood pressure) and cancer growth (tumor, localized spread, blood test results).*



	Blood pressure	Heart rate	Respiratory rate	Platelets	Lymphocites	Red cells	BMI	survival	age	Body fat	cholesterol	.... + 20000 genes expression
Patient 1	....	....	....	....	....	....	....	....	....	....	....	... ...
Patient 2	....	....	....	....	....	....	....	....	....	....	....	... ...
Patient 3	....	....	....	....	....	....	....	....	....	....	....	... ...
Patient 4	....	....	....	....	....	....	....	....	....	....	....	... ...
Patient 5	....	....	....	....	....	....	....	....	....	....	....	... ...
Patient 6	....	....	....	....	....	....	....	....	....	....	....	... ...
Patient 7	....	....	....	....	....	....	....	....	....	....	....	... ...
Patient 8	....	....	....	....	....	....	....	....	....	....	....	... ...
Patient 9	....	....	....	....	....	....	....	....	....	....	....	... ...
....	....	....	....	....	....	....	....	....	....	....	....	... ...
....	....	....	....	....	....	....	....	....	....	....	....	... ...
....	....	....	....	....	....	....	....	....	....	....	....	... ...
....	....	....	....	....	....	....	....	....	....	....	....	... ...
....	....	....	....	....	....	....	....	....	....	....	....	... ...
....	....	....	....	....	....	....	....	....	....	....	....	... ...

FIGURE 1. **an overview of an high-dimensional dataset with P=20011 features and n=200 observations**

Examples of applications, including in social science, are extremely numerous; see **Plomin (2018)**.

A. *Challenge when Analysing High-dimensional Data*

Analyses of high-dimensional data require consideration of potential problems that come with having more features than observations. Such datasets pose a challenge for data analysis as standard methods of analysis, such as *least squares linear regression*, are no longer appropriate. Many of the issues that arise in the analysis of high-dimensional data are known in classical approaches, since they apply also when $n > p$:

these include the role *bias-variance trade-off* and the danger of *over-fitting*. Though these issues are always relevant, they can become particularly important when the number of features is very large relative to the number of observations.

In order to illustrate the need for extra care and specialized techniques for regression when $p > n$, we begin by examining what can go wrong if we apply a statistical technique not intended for high-dimensional settings. For this purpose, we examine *least squares regression*. But the same concepts apply to *logistic regression, linear discriminant analysis*, and other classical statistical approaches.

1) Setup of Linear Regression Model

The general form of the multiple linear regression model is as follows:

$$Y = \mathbb{E}[Y|X] + \epsilon \;=\; \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \epsilon \qquad (1)$$

Where y is the dependent variable, $\beta_0, \beta_1, ..., \beta_p$ are regressions coefficients, and $X_1, ..., X_p$ are independents variables in the model; $\mathbb{E}[Y]$ the expectation of the response variable. In the classical regression setting, it is usually assumed that the error term ϵ follows the *normal distribution* with mean $\mathbb{E}[\epsilon] = 0$ and constant variance $Var[\epsilon] = \sigma^2$.

We consider a datasets from the following model

$$Y_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip} + \epsilon_i, \; i = 1, ..., n \qquad (2)$$

Where X_{ij} is the j^{th} variable for individual i and $\epsilon_i' s$ are random errors assuming $\mathbb{E}[\epsilon_i] = 0$ and $Var[\epsilon_i|X] = \sigma^2 \; for \; i = 1, 2, ..., n$. The data from this model can be written in matrix form:

$$y = X\beta + \epsilon, \qquad (3)$$

where:

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \; X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \; \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}$$

$\epsilon = (\epsilon_1, ..., \epsilon_n)^T$

The regression parameter are estimated by minimizing ordinary least squares:

$$(y - X\beta)^t (y - X\beta) = \| y - X\beta \|^2, \; (\| \, . \, \|^{1}).$$

2) Ordinary Least Squares Estimates (OLS Estimates)

Proposition II.1 (*from [25]*). *The least squares estimation of β for the linear regression model is given by,*

$$b = argmin_\beta \left\{ \; \| y - X\beta \|_2^2 \; \right\} = (X^t X)^{-1} X^t y, \qquad (4)$$

assuming $(X^t X)$ is a non-singular matrix.Note that this is equivalent to assuming that the matrix X is of full rank[2].

The estimator $b = (X^t X)^{-1} X^t y$ is and unbiased estimator of β. In addition,its covariance matrix is given by,

$$Cov(b) = (X^t X)^{-1} \sigma^2. \qquad (5)$$

Proposition II.2 (*from [25]*). *The unbiased estimator of the variance σ^2 in the multiple linear regression is given by:*

$$s^2 = \frac{RSS}{n - p - 1} = \frac{1}{n - p - 1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \qquad (6)$$

The proof is straightforward using the following lemmas:

Lemma 1. *Let $A_{n \times n}$ be and idempotent matrix of rank p then the eigenvalues of A are either* 1 *or* 0.

Lemma 2. *If A is and idempotent matrix, then $tr(A) = rank(A) = p$.*

Lemma 3. *Let $y^t = (y_1, y_2, ..., y_n)$ be an $n \times 1$ vector with mean $\mu^t = (\mu_1, ..., \mu_n)$ and variance σ^2 for each component. Further, it is assumed that $y_1, y_2, ..., y_n$ are independent.Let A be and $n \times n$ matrix.*

The expectation of the quadratic form of random variables is given by:

$$\mathbb{E}[y^t A y] = \sigma^2 tr(A) + \mu^t A \mu, \qquad (7)$$

Now that a brief presentation of the linear model has been made,let's come back to the main question to know the problems encountered in high-dimension settings.

- **Theoretically:** *when $p > n$,* $X^t X$ is not invertible (or near singular) and s^2 in (6) is not defined.

Lemma 4. *An $n \times n$ ill-conditioned or near singular matrix has at least one of its eigenvalues close to zero, and then the eigenvalue of the inverse tend to be very large.*

Proposition II.3 (*from [25]*). *The average Euclidean distance measure $\mathbb{E}[\| \; b - \beta \; \|^2]$ between the least squares estimate b and the true parameter β is given by:*

$$\mathbb{E}[\| \; b - \beta \; \|^2] = \sigma^2 tr[(X^t X)^{-1}] \qquad (8)$$

Remark II.1. *Assuming that $(X^t X)$ has k distinct eigenvalues $\lambda_1, ..., \lambda_k$, then the eigenvalues of $(X^t X)^{-1}$ are $\frac{1}{\lambda_1}, ..., \frac{1}{\lambda_k}$, denoting by $V = (v_1, ..., v_k)^t$ the corresponding normalized eigenvectors, we can write $V^t (X^t X)^{-1} V = D = diag(\frac{1}{\lambda_1}, ..., \frac{1}{\lambda_k})$.*
Moreover, $tr(X^t X)^{-1}) = tr(V^t V (X^t X)^{-1}) = tr(V^t (X^t X)^{-1} V) = tr(D) = \sum_{i=1}^{k} \frac{1}{\lambda_i}$; we then have:

$$\mathbb{E}[\| \; b - \beta \; \|^2] = \sigma^2 \sum_{i=1}^{k} \frac{1}{\lambda_i} \; \Leftrightarrow \; \mathbb{E}[\| \; b \; \|^2] = \| \; \beta \; \|^2 + \sigma^2 \sum_{i=1}^{k} \frac{1}{\lambda_i}. \qquad (9)$$

Now it is easy to see that if one of $\lambda_i, \; i = 1, ..., k$ is very small, say for instance $\lambda_i = 0.00001$ then roughly, $\| \; b \; \|^2 = \sum_{i=1}^{k} b_i^2$ may **over estimate** $\| \; \beta \; \|^2 = \sum_{i=1}^{k} \beta_i^2$ by $10000\sigma^2$ times.

The above discussions indicate that if some columns in X are highly correlated with other columns in X, then from *lemma(4)*, the covariance matrix $Cov(b) = (X^t X)^{-1} \sigma^2$ will have one or more large eigenvalues, so the mean Euclidean distance of $\mathbb{E}[\| \; b - \beta \; \|^2]$ will be inflated. Consequently, this makes the estimation of the regression parameter β less reliable. Thus, the high levels of correlation between variables in high-dimensional datasets will have a negative impact on the least square estimates of the regression parameter.

It is evident that alternative approaches that are more suitable for the high-dimensional setting are necessary.

[1] $\| \, . \, \|$ is the Euclidean norm on $\mathbb{R}^n$

[2] i.e, $rank(X) = p + 1 < n$, this then implies that $rank(X^t X) = p + 1$ and therefore that $X^t X$ is invertible.

III. Some Appropriate Statistical Methods for Handling High-Dimensional Data

A. Ridge Regression

Ridge regression is one of the remedial measures for handling severe multicollinearity in least squares estimation. Multicollinearity occurs when the predictors included in the linear model are highly correlated with each other. When this is the case, the matrix X^tX tends to be singular or ill-conditioned and hence identifying the least squares estimates will encounter numerical problems.

Proposition III.1. $\mathbb{E}[\| b-\beta \|^2] = \sum_{j=1}^{n}(\mathbb{E}[b_j]-\beta_j)^2 + \sum_{j=1}^{n} Var[b_j]$

According to the "Gauss-Markov" theorem, the least squares approach achieves the smallest variance among all unbiased linear estimates. This, however, does not necessarily guarantee the minimum **MSE**.

To better distinguish different types of estimators, let $\hat{\beta}^{LS}$ denote the ordinary least square estimator of β. We have shown that $MSE(\hat{\beta}^{LS}) = \mathbb{E}[\| \hat{\beta}^{LS}-\beta \|^2] = \sigma^2 tr[(X^tX)^{-1}]$ (8). Thus,

$\mathbb{E}[\| \hat{\beta}^{LS} \|^2] = \| \beta \|^2 + \sigma^2 tr[(X^tX)^{-1}]$ (9); it can be seen that, with ill-conditioned X^tX, the resultant LSE $\hat{\beta}^{LS}$ would be large in length $\| \hat{\beta}^{LS} \|$ and associated with inflated standard error (see (9)). This inflated variation would lead to poor model prediction as well.

The Ridge regression is a constrained version of least squares. It tackles the estimation problem by providing a biased estimator yet with small variance.

Theorem III.1. *For any estimator b, the least squares criterion $\mathcal{Q}(b) = \| y-Xb \|^2$ can be rewritten as its minimum, reached at $\hat{\beta}^{LS}$ plus a quadratic form in b.*

$$\begin{aligned}\mathcal{Q}(b) &= \| y-Xb \|^2 \\ &= \underbrace{\| y-X\hat{\beta}^{LS} \|^2}_{\mathcal{Q}_{min}} + \underbrace{(\hat{\beta}^{LS}-b)^tX^tX(\hat{\beta}^{LS}-b)}_{\phi(b)} \\ &= \mathcal{Q}_{min} + \phi(b)\end{aligned} \tag{10}$$

The contour for each constant of the quadratic form $\phi(b)$ are hyper-ellipsoids centred at the ordinary LSE $\hat{\beta}^{LS}$.

The optimization problem in Ridge regression can be stated as:

$$minimize \ \| \beta \|^2 \ \ S.t \ (\hat{\beta}^{LS}-b)^tX^tX(\hat{\beta}^{LS}-b) = \phi_0.$$

$(\phi_0 \ \ constant)$

The enforced constraint guarantees a relatively small residual sum of squares $\mathcal{Q}(\beta)$ when compared to its minimum $\mathcal{Q}_{min}$. As a Lagrangian problem, it is equivalent to

$$minimize \ f(\beta) = \| \beta \|^2 + \frac{1}{k}[(\hat{\beta}^{LS}-b)^tX^tX(\hat{\beta}^{LS}-b) - \phi_0],$$

Where $\frac{1}{k}$, $(k > 0\ )$ is the multiplier chosen to satisfy the constraint.

Proposition III.2 (Hoerl and Kennard (1970))**.** *The numerical solution of this problem corresponding to the Ridge regression estimator of β is,*

$$\hat{\beta}^R = (X^tX + k\mathbb{I}_p)^{-1}X^ty \tag{11}$$

An equivalent way is to write the Ridge problem in the penalized or constrained least squares form by:

$$minimize \ \| y-X\beta \|^2 \quad S.t \ \| \beta \|^2 \leq s \ for \ some \ constant \ s \tag{12}$$

The Lagrangian problem becomes

$$minimize \ \ \| y-X\beta \|^2 + \lambda \| \beta \|^2 \tag{13}$$

which yields the same estimator given in (11). The penalty parameter $\lambda \geq 0$ controls the amount of shrinkage in $\| \beta \|^2$. The larger the value of λ, the greater the amount of shrinkage. For this reason, the Ridge estimator is also called the shrinkage estimator. There is a one-to-one correspondence among λ, s, k, and ϕ_0.

Let $\lambda_{max} = \lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p = \lambda_{min}$ denote the eigenvalues of X^tX, then the corresponding eigenvalues of Z are $\frac{\lambda_j}{\lambda_j+k}$, $j = 1, ..., p$. From (9), $MSE(\hat{\beta}^{LS}) = \sigma^2\sum_{j=1}\frac{1}{\lambda_j}$.

Proposition III.3. *If $MSE(\hat{\beta}^R, k)$ denote the mean square error of ridge regression estimator, then*

$$\begin{aligned}MSE(\hat{\beta}^R, k) &= k^2\beta^t(X^tX + k\mathbb{I})^{-2}\beta + \sigma^2\sum_j\frac{1}{\lambda_j}.\frac{\lambda_j^2}{(\lambda_j+k)^2} \\ &= \lambda_1(k) + \lambda_2(k).\end{aligned} \tag{14}$$

Theorem III.2 (*Hoerl and Kennard (1970)*)**.** *There always exists a $k > 0$ such that,*

$$MSE(\hat{\beta}^R, k) < MSE(\hat{\beta}^R, 0) = MSE(\hat{\beta}^{LS})$$

B. Lasso Regression

The Lasso (Least Absolute Shrinkage and Selection Operator) is another shrinkage method like Ridge regression, yet with an important and attractive feature in variable selection.

Ridge regression does have one obvious disadvantage; unlike *best subset, forward step-wise, backward step-wise*[3], which will generally select models that involve just a subset of variables, Ridge regression will include all p predictors in the final model. The penality $\lambda \| \beta \|^2$

[3] methods used in low-dimension regression to select the most appropriate variables for a best model

in (13) will shrink all of the coefficients towards zero, but it will not set any of them exactly to zero (unless $\lambda = \infty$). This may not be a problem for prediction accuracy, but it can create a challenge in model interpretation in settings in which the number of variables p is quite large. Increasing the value of λ will tend to reduce the magnitudes of the coefficients, but will not result in exclusion of any of the variables.

The Lasso is a relatively recent alternative to Ridge regression that overcomes this disadvantage . The Lasso estimator of β is obtained by

$$minimizing \left\{ \parallel y - X\beta \parallel_2^2 \right\} \quad S.t \sum_{j=1}^{p} |\beta_j| \leq s \qquad (15)$$

for some constant s.
Namely, the L_2 penality $\parallel \beta \parallel^2 = \sum_{j=1}^{p} \beta_j^2$ in Ridge regression is replaced by the L_1 penality $\parallel \beta \parallel_1 = \sum_{j=1}^{p} |\beta_j|$ in Lasso. The Lagrangian problem becomes:

$$minimize_{\beta \in \mathbb{R}^p} \{ \parallel y - X\beta \parallel^2 + \lambda \parallel \beta \parallel_1 \}. \qquad (16)$$

C. Lasso Regression

The Lasso (Least Absolute Shrinkage and Selection Operator) is another shrinkage method like Ridge regression, yet with an important and attractive feature in variable selection.

Ridge regression does have one obvious disadvantage; unlike *best subset, forward step-wise, backward step-wise*[4], which will generally select models that involve just a subset of variables, Ridge regression will include all p predictors in the final model. The penality $\lambda \parallel \beta \parallel^2$ in (13) will shrink all of the coefficients towards zero, but it will not set any of them exactly to zero (unless $\lambda = \infty$). This may not be a problem for prediction accuracy, but it can create a challenge in model interpretation in settings in which the number of variables p is quite large. Increasing the value of λ will tend to reduce the magnitudes of the coefficients, but will not result in exclusion of any of the variables.

The Lasso is a relatively recent alternative to Ridge regression that overcomes this disadvantage . The Lasso estimator of β is obtained by

$$minimizing \left\{ \parallel y - X\beta \parallel_2^2 \right\} \quad S.t \sum_{j=1}^{p} |\beta_j| \leq s \; (s\; constant\;) \qquad (17)$$

Namely, the L_2 penality $\parallel \beta \parallel^2 = \sum_{j=1}^{p} \beta_j^2$ in Ridge regression is replaced by the L_1 penality $\parallel \beta \parallel_1 = \sum_{j=1}^{p} |\beta_j|$ in Lasso. The Lagrangian problem becomes:

$$minimize_{\beta \in \mathbb{R}^p} \{ \parallel y - X\beta \parallel^2 + \lambda \parallel \beta \parallel_1 \}. \qquad (18)$$

1) Computation of Lasso Solution

The Lasso problem is a convex program, specifically a quadratic program (**QP**) (*visit [14] for more detail*). with a convex constraint. As such, there are many sophisticated **QP** methods for solving the Lasso. However, there is a particularly simple and effective computational algorithm that gives insight into how the Lasso works. The Lagrangian form (18) is especially convenient for numerical computation of the solution.

2) Theoretical properties of Lasso penalty

A common assumption of the Lasso model is **sparsity**, i.e., only a small number of covariates influence the outcome.

Let $S = \{j : \beta_j \neq 0\}$ be the index set of non-zero components of the true coefficient vector $\beta \in \mathbb{R}$ and denote the number of relevant covariates by $s = card\{S\}$. Under the sparsity assumption, most components of β are zero such that $s \ll p$. For any $\lambda \geqslant 0$, define the active set of the Lasso as $\hat{S}(\lambda) = \{j : \hat{\beta}_j(\lambda) \neq 0\}$. Given β, we order the covariates such that $S = \{1, \ldots, s\}$, $S^c = \{s+1, \ldots, p\}$ and consider the partitioning $X = (X_S, X_{S^c})$ where $X_S \in \mathbb{R}^{n \times s}$ contains the n measurements of the s relevant covariates, and $X_{S^c} \in \mathbb{R}^{n \times (p-s)}$ contains the n measurements of the $(p-s)$ irrelevant covariates. Sample covariance matrices are denoted by Σ_X and the empirical covariance is given by $S_{XX} = \frac{X^T X}{n}$.

State the following basic inequality,

Lemma 5 ([1],P.103). *we have,*

$$\frac{1}{n} \parallel X\hat{\beta}^{Lasso} - X\beta \parallel_2^2 + \lambda \parallel \hat{\beta}^{Lasso} \parallel_1 \leq \frac{2\epsilon^t X(\hat{\beta}^{Lasso} - \beta)}{n} + \lambda \parallel \beta \parallel_1 \qquad (19)$$

Now let us introduce the set,

$$\mathcal{A} = \{ \frac{2}{n} \parallel \epsilon^t X \parallel_\infty \leq \lambda_o \},$$

for a suitable value of λ_o, the set $\mathcal{A}$ has large probability. Indeed, with Gaussian errors this follow from the following lemma:

Lemma 6 ([1], P.104). *Suppose that the diagonal elements of the Gram matrix $\frac{X^T X}{n}$ equal 1 for all j.Then we have for all $t > 0$ and for $\lambda_0 = 2\sigma\sqrt{\frac{t^2 + 2\log(p)}{n}}$,*

$$\mathbb{P}(\mathcal{A}) \geq 1 - 2\exp(-\frac{t^2}{2}) \qquad (20)$$

Corollary 1 (Lasso estimation consistency). *Let the assumption of lemma **6** hold. For some $t > 0$, let the regularization parameter be $\lambda = 2\hat{\sigma}\sqrt{\frac{t^2 + 2\log(p)}{n}}$, where $\hat{\sigma}$ is some estimator of σ. Then with probability at least $1 - \alpha$, where $\alpha = 2\exp(-\frac{t^2}{2}) + \mathbb{P}(\hat{\sigma} \leq \sigma)$.We have:*

[4] methods used in low-dimension regression to select the most appropriate variables for a best model

$$\frac{2}{n} \parallel X(\hat{\beta}^{Lasso} - \beta) \parallel_2^2 \leq 3\lambda \parallel \beta \parallel_1 \tag{21}$$

we thus conclude that, taking the regularisation parameter λ of order $\sqrt{\frac{\log(p)}{n}}$ and assume that $\parallel \beta \parallel_1 = o\big(\sqrt{\frac{n}{\log(p)}}\big)$, result in consistency of the Lasso.

This means that , up to the $\log(p) - term$ and compatibility constant Φ_o^2, the mean squared prediction error is of the same order as if one knew a priori which of the covariates are relevant and using ordinary least squares estimation based on the true relevant s only.

See also [*Theorem 14.6, Chap 14 from Guedon et al. (2007)*] for the corresponding result for the random design.

Let us define the vectors β_S and β_{S^c} by:

$$\beta_{j,S} = \beta_j \mathbb{1}_{\{j \in S\}}, \quad \beta_{j,S^c} = \beta_j \mathbb{1}_{\{j \notin S\}}. \tag{22}$$

Clearly, $\beta = \beta_S + \beta_{S^c}$; β_S has zeroes outside the index set S and the elements of β_{S^c} can only be non-zero in the complement S^c of S.

Definition III.1 (Compatibility condition)**.** *We say the the compatibility conditionis met for the set S if for some $\Phi_0 > 0$ and for all $\beta \in \mathbb{R}^p$ such that $\parallel \beta_{S^c} \parallel_1 \leq 3 \parallel \beta_S \parallel_1$, it holds that*

$$\parallel \beta_S \parallel_1^2 \leq \frac{1}{n} \frac{s \parallel X\beta \parallel_2^2}{\Phi_o 2^2} = \frac{s(\beta^t S_{XX} \beta)}{\Phi_o^2}. \tag{23}$$

Theorem III.3 (*[1], Theorem 6.1,P.107*)**.** *Suppose the compatibility condition holds for S.Then on $\mathcal{A}$, we have for $\lambda \geq 2\lambda_0$*

$$\frac{1}{n} \parallel X(\hat{\beta}^{Lasso} - \beta) \parallel_2^2 + \lambda \parallel \hat{\beta}^{Lasso} - \beta \parallel_1 \leq \frac{4\lambda^2 s}{\Phi_o^2}. \tag{24}$$

Lemma 7. *On $\mathcal{A}$, with $\lambda \geq 2\lambda_0$ we have:*

$$\frac{2}{n} \parallel X(\hat{\beta}^{Lasso} - \beta) \parallel_2^2 + \lambda \parallel \hat{\beta}_{S^c}^{Lasso} \parallel_1 \leq 3\lambda \parallel \hat{\beta}_S^{Lasso} - \beta_S \parallel_1 . \tag{25}$$

Remark III.1. *The theorem combines two results:*

$$\frac{2}{n} \parallel X(\hat{\beta}^{Lasso} - \beta) \parallel_2^2 \leq \frac{4\lambda^2 s}{\Phi_o^2}, \ (prediction\ error\ bound) \tag{26}$$

$$\parallel \hat{\beta}^{Lasso} - \beta \parallel_1 \leq \frac{4\lambda^2 s}{\Phi_o^2}, \ (L_1 - error\ bound\ ) \tag{27}$$

Corollary 2 (estimation accuracy of β,[*Knight and Fu (2000)*)**.** *Under compatibility assumptions on design matrix X and on the sparsity $s = card\{S\}$, for lambda in the suitable range of order $\lambda \approx \sqrt{\frac{log(p)}{n}}$,*

$$\parallel \hat{\beta}^{Lasso} - \beta \parallel_1 \xrightarrow[n \to +\infty]{\mathbb{P}} 0 \ ; \quad \parallel \hat{\beta}^{Lasso} - \beta \parallel_2 \xrightarrow[n \to +\infty]{\mathbb{P}} 0 \tag{28}$$

Knowing that Lasso is widely use for model selection, it is necessary to assess how well the sparse model given by Lasso relates to the true model.We make this assessment by investigating Lasso's model consistency (under linear model); That is, for $S = \{j, \beta_j \neq 0\}$ being the true active set, we look for a Lasso procedure delivering an estimator $\hat{S} = \{j, \hat{\beta}_j^{Lasso} \neq 0\}$ of S such that $\hat{S} = S$ with large probability.

Since using Lasso estimate involves choosing the appropriate amount of regularization, to study the model selection consistency of the Lasso, we consider two problems: whether there exists a deterministic amount of regularization that gives consistent selection, or for each random realization whether there exists a correct amount of regularization that selects the true model.The so-called **"irrepresentable condition"** thoroughly interpreted by *Zhao and Yu (2006) [26]* is almost necessary and sufficient for both types of consistency.

An estimate which is consistent in term of parameter estimation does not necessarily consistently select the correct model (or even attempt to do so) where the reverse is also true.The former requires $\hat{\beta}^{Lasso} - \beta \xrightarrow[n \to +\infty]{\mathbb{P}} 0$ while the latter requires $\mathbb{P}\big(\{\hat{S} = S\}\big) \xrightarrow[n \to +\infty]{\mathbb{P}} 1$. We desire our estimate to have both consistencies. However, to separate the selection aspect of consistency from the parameter estimation aspect.We make the following definitions about *"sign[5] consistency"* that does not assume the estimates to be estimation consistent.

Definition III.2. *An estimate $\hat{\beta}_n$ is equal in sign with the true model β if and only if,*

$$Sign(\hat{\beta}_n) = Sign(\beta)$$

Definition III.3. *Lasso is strongly sign consistent if there exists $\lambda_n = f(n)$, that is , a function independent of Y and X such that:*

$$\lim_{n \to \infty} \mathbb{P}\big(\{Sign(\hat{\beta}^{Lasso}) = Sign(\beta)\}\big) = 1, \ (*)$$

Definition III.4. *Lasso is general sign consistentcy if*

$$\mathbb{P}\big(\{\exists \lambda \geq 0, Sign(\hat{\beta}^{Lasso}) = Sign(\beta)\}\big) = 1, \ (**)$$

Remark III.2.
- *Strong sign consistency implies one can use a preselected λ to achieve consistent model selection via Lasso.*
- *General sign consistency means for a random realization there exists a correct amount of regularization that select true model.*
- $(*) \Rightarrow (**)$

[5] $Sign(.)$ maps positive entry to 1 ,negative to -1 and 0 to 0

Definition III.5 (Irrepresentable Condition)**.** *We say that , Irrepresentable condition is met for the set S if there exists a constant $\theta \in [0,1[$ such that ,*

$$\| S_{XX}(S^c,S)S_{XX}(S,S)^{-1}sign(\beta_S) \|_\infty \leq \theta.. \quad (29)$$

Theorem III.4 (Variables selection consistency,[*Zao and Yu (2006)*)**.** *]*

The irrepresentable condition (29) *for the active set S is a sufficient and essentially necessary condition for Lasso to select only variables in active set S; that is to achieve sign consistency.*

Proof. refer to *Zao and Yu (2006) [26]* or *Meinshausen and Buhlmann (2010)* for more details.

Remark III.3. *The irrepresentable condition , as given in* (29) *depends on the Gram matrix $\frac{X^tX}{n}$ but also on the signs of the true unknown parameter β, whereas the compatibility condition* (23) *only depends on Σ_X.*

D. Dantzig Selector (DS)

The Lasso is not the only L_1-penalization possible. From the score equation, the Dantzig Selector by *Candes and Tao [3]* also belongs to the class of regularization methods in regression. It can be formulated as the Lasso but instead of controlling the squared error loss, it controls the correlation of residuals with X. Specifically, the Dantzig selector estimator is defined to be the solution of the minimization problem:

$$\begin{aligned}&\underset{\beta\in\mathbb{R}^p}{min}\{ \ \| \beta \|_1 \ \} \\ &subject\ to \ \ \| X^t(y - X\beta) \|_\infty := \underset{1\leq i \geq p}{sup} |(X^tr)_i| \leq \lambda_p.\sigma,\end{aligned} \quad (30)$$

for some $\lambda_p > 0$, where $r = y - X\beta$ is the residual vector.

Remark III.4. *The constraint on the residual vector imposes that for each $j \in \{1,\ldots,p\}$, $|(X^tr)_j| \leq \lambda_p\sigma$, which guarantees that the residuals are within the noise level.*

The Dantzig selector and Lasso are closely related. Connections between the Dantzig Selector and the Lasso have been discussed in *Jame et al. (2008)* where it is shown that under some general conditions, the Dantzig Selector and the Lasso produce the same solution path.

Both models share the feature of setting some of the parameters to zero, i.e. they perform variable selection.

Remark III.5. *Though under some general conditions, the Lasso and Dantzig may produce the same solution path, they differ conceptually in that the Dantzig stems directly from an estimating equation, whereas the Lasso stems from a likelihood or an objective function.*

The theoretical results (estimation accuracy and model selection consistency) for the Dantzig selector estimator are provided with detailed supporting proof in [*[3], theorem 1.1; theorem 1.2*]

E. Elastic-Net Regression

We ended the section on Lasso regression by saying that it works best when your model contains a lot of useless variables. We also said that Ridge regression works best when most of the variables in your model are useful.

Remark III.6. *When we know about all of the parameters in our model, it's easy to choose if we want to use Lasso or Ridge regression; but what do we do when we are in a high dimension setting where the model includes tons more variables, far too many to know everything about?*

When you have millions of parameters, then you will almost certainly need to use some sort of regularization to estimate them. However, the variables in those models might be useful or useless; we do not know in advance. So how do we choose if we should use Lasso or Ridge regression?The good news is that we don't have to choose, instead, we use *Elastic-Net* regression. Just like Lasso and Ridge regression, Elastic-Net regression starts with least squares. Then, it combines the Lasso regression penalty $\lambda_1 \| \beta \|_1$ with the Ridge regression penalty $\lambda_2 \| \beta \|_2^2$. The Lagrangian problem becomes

$$\text{minimize}_\beta\{\| y - X\beta \|^2 + \lambda_1 \| \beta \|_1 + \lambda_2 \| \beta \|^2\}$$

Altogether, Elastic-Net regression combines the strengths of Lasso and Ridge regression. Note that the Lasso and Ridge regression penalties get their own λ's: λ_1 for Lasso and λ_2 for Ridge. But more often, the problem is written as

$$\textbf{minimize}_\beta\{\| y - X\beta \|^2 + \lambda(\alpha \| \beta \|_1 + (1-\alpha) \| \beta \|^2)\}, \quad \textbf{for } \alpha \in [0,1] \text{ and } \lambda \geq 0\text{,say}$$

$$\hat{\beta}^E(\lambda,\alpha) = \text{argmin}_\beta\{\| y - X\beta \|^2 + \lambda(\alpha \| \beta \|_1 + (1-\alpha) \| \beta \|^2)\}$$

We still have the regularization parameter λ, but we only have one regularization parameter common to both terms. We also have a parameter α which controls the mix between L_1 and L_2 regularization.

Remark III.7. *We notice that:*

- $\hat{\beta}^E(\lambda,1) = \hat{\beta}^{Lasso}(\lambda)$,
  $\hat{\beta}^E(\lambda,0) = \hat{\beta}^R(\lambda)$, $\quad \hat{\beta}^E(0,\alpha) = \hat{\beta}^{LS}$
- *and when $\alpha \notin \{0,1\}$ and $\lambda \neq 0$, we obtain a hybrid of Ridge and Lasso estimation.*

1) Cross-validation to find the best value of λ

There are various methods to select the "best" value for λ. One method is to split the data into **K** chunks. We then use **K-1** of these as a training set, and the remaining 1 chunk as the test set. We can repeat this until we've rotated through all **K** chunks, giving us a good estimate of how well each of the lambda values work in our data. This is called *cross-validation*, and doing this repeated *test/train* split gives us a better estimate of how generalizable our model is.

We can use this new idea to choose a lambda value by finding the lambda that minimizes the error across each of the test and training splits.

Let (X_k, y_k) denote the subset of X and y for the k-th fold, with $k = 1, \dots, K$. The optimal λ is obtained by minimizing the total *Cross-validation* error:

$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmin}} \underbrace{\left\{ \frac{1}{K} \sum_{k=1}^{K} \frac{1}{n_k} \parallel y_k - X_k \hat{\beta}_k(\lambda) \parallel_2^2 \right\}}_{CV_{(K)}}, \tag{31}$$

IV. Numerical Implementation

We present here two illustrative numerical applications. The first one is based on simulated data and the last one on real data. The purpose of the numerical experiment is to show the behavior and to investigate if there was a difference in predictive power between the previous three regularization methods: ridge, Lasso, and Elastic-net regression when they were applied to high-dimensional data. The statistical analysis was implemented using **R** statistical software.

A. Simulated Data

For the simulation study, we used generalized linear models (GLMs) for penalized logistic regression. The *"glmnet"* [13] package for **R** fits a GLM via penalized maximum likelihood. We will not provide a theory about GLMs in this study; for specific information regarding GLMs, we refer to [18]. The measures that are used to assess how good a logistic regression model is for prediction are the *misclassification error rate* (ME), which denotes the fraction of incorrect classifications over all observations, and the *Area Under Curve* (AUC), which is a measure of discrimination ranging values between 0 and 1 (visit [16] for more details). The simulation study was inspired by the paper by *Krona* [16]. However, adjustments were made to the simulated datasets.

a: Process description:

The simulated data consisted of four independent high-dimensional datasets. Each dataset was divided into a training and a test set. The three methods were used to fit a corresponding model to each of the training sets. The fitted models were used to make predictions for each of the corresponding test sets. Finally, we computed the AUC, the ME, and extracted the number of non-zero $\hat{\beta}$-coefficients. The procedure was repeated 100 times per example.

b: Simulation design:

We simulated $p = 1000$ predictors and $n = 200$ observations such that $p >> n$ and the data qualified as high-dimensional. All predictor variables X were continuous multivariate normally distributed, except for the binary response variable Y. A multiple group of predictors with varying strength of correlation was simulated for each dataset. The predictors were generated by sampling from a multivariate normal distribution with the following probability density function:

$$f_X(x) = \frac{1}{(2\pi)^{\frac{n}{2}} \sqrt{\det(\Sigma)}} \exp\left\{ -\frac{1}{2}(x-\mu)^t \Sigma^{-1} (x-\mu) \right\}$$

where μ is the mean vector and $\Sigma = (\rho_{ij})_{i,j}$ is the covariance matrix. For all x, we set $\mu = 0$ and $Var[x] = 1$. Thus, Σ equals the correlation matrix of X. Each predictor variable was assigned a predetermined β-value. The response variable was simulated by running the simulated data through the inverse logit function (see [18]),

$$\pi(x) = \frac{1}{1 + e^{-X^t \beta}}.$$

Given the threshold $\pi_0 = 0.5$, the observed value was categorized into one of the two classes: $Y = 1$ if $\pi(x) > 0.5$ and $Y = 0$ if $\pi(x) \leq 0.5$. Consequently, we obtained a vector Y and a matrix X consisting of 200 observations of the binary response variable and the predictor variables, respectively.

Detailed information about the four examples:

- **Example 1:** We set the pairwise correlation between predictors X_i and X_j to $\rho_{ij} = 0.5^{|i-j|}$. We assigned the first 122 β-coefficients a specified vector that consisted of random values within [2,5]. The remaining coefficients were set to 0.
- **Example 2:**: We set $\rho_{ij} = 0.5^{|i-j|}$. We set all coefficients to be $\beta = 0.8$.
- **Example 3:** We set $\rho_{ij} = 0.9^{|i-j|}$. The coefficients were split into 8 groups, where the coefficients were pairwise set to 0 and 2: $\beta = (\underbrace{2,2,\dots,2}_{125}, \underbrace{0,0,\dots,0}_{125}, \underbrace{2,2,\dots,2}_{125}, \dots)^t$.
- **Example 4:** The pairwise correlation between the first 500 predictors X_i and X_j $(1 \leq i, j \leq 500)$ were set to $\rho_{ij} = 0.5^{|i-j|}$, and the pairwise correlation for the remaining predictors were set to 0. We set the first 500 coefficients to $\beta = 3$ and the remaining coefficients to 0: $\beta = (\underbrace{3,3,\dots,3}_{500}, \underbrace{0,0,\dots,0}_{500})^t$.

c: Results :

for each simulation, we calculated AUC, ME, and their standard deviations (sd). Additionally, we calculated the average number of selected variables by Lasso and the Elastic net. The results are summarized in Table 1.

In Example 1, a small subset of predictors were assigned non-zero β-coefficients. On average, Lasso and the Elastic net selected 28 and 316 variables, respectively. We observe that Ridge regression has the highest AUC and the lowest ME.

In Example 2, the predictors were assigned coefficients $\beta = 0.8$ with relatively high correlation among predictors. As demonstrated in Table 1, Ridge regression improves over other methods in terms of AUC and ME. As mentioned in Subsection 1.3.1, Ridge regression tends to perform well under the circumstances in Example 2. Furthermore, the average number of coefficients for Lasso and Elastic net was 20 and 328, respectively. In this setting, the Elastic net identifies a larger number of coefficients that are correlated and non-zero. On the other hand, Lasso results in a sparse final model but identifies fewer non-zero coefficients. Instead, the chosen model resulted in a high ME (Table 1).

In Example 3, the predictors were divided into 8 groups and pairwise assigned coefficients of 0 and 2. We observe that Ridge regression outperforms Lasso and Elastic net in terms of AUC. Since Elastic net and Ridge regression perform similarly, they seem to perform equally well in this setting. As discussed earlier, Ridge regression includes all predictors in the final model and results in a less interpretable model. However, Elastic net identifies, on average, 415 non-zero coefficients. Presumably, Elastic net adopts the grouping effect and correctly identifies almost all non-zero coefficients simultaneously, achieving high prediction accuracy.

In Example 4, the predictors were divided into two groups of equal size, assigned $\beta = 3$ and $\beta = 0$, respectively. The first 500 were correlated, while the remaining 500 predictors were uncorrelated. As seen in Table 1, Ridge regression achieved the highest AUC, while Elastic net managed to identify approximately all non-zero coefficients as a result of the grouping effect.

d: Summary

The results show that the three methods perform well in the sense that AUC $\geq$ 0.5 in examples 1-4. We observe that despite the fact that ridge regression tends to spread the coefficient shrinkage over a large number of coefficients, it achieves high predictive power throughout examples 1-4. Especially, the results in example 3 demonstrate the capacity of ridge regression. We identify that when the number of predictors is very large and a larger fraction of them must be included in the model, ridge regression dominates the Lasso and the elastic net. Consequently, it confirms that ridge regression is a satisfactory method for prediction on correlated datasets.

The results from example 2 determine that the Lasso is outperformed by the elastic net. Furthermore, we observed that the elastic net benefits from the ability to put a larger weight on the quadratic penalty, while simultaneously shrinking some coefficients to zero by the absolute penalty. Moreover, we observe that ridge regression and the elastic net generally improve over the Lasso. We can see that the elastic net approximately identifies all non-zero coefficients in the simulations. In example 4, the elastic net performed grouped selection and proved to be a better variable selection method than the Lasso.

Even though ridge regression did not incorporate variable selection, it achieved high prediction accuracy throughout examples 1-4. Therefore, we observe that if interpretability is not fundamental, ridge regression manages to accomplish high predictive power. Ultimately, the elastic net has the advantage of incorporating variable selection. Consequently, its final model is more interpretable than that of ridge regression.

B. Real data example

a: Data description:

For the real data example, we will be working with *"human DNA methylation data"* from *"flow-sorted blood samples"*. DNA methylation assays measure, for each of many sites in the genome, the proportion of DNA that carries a methyl mark (a chemical modification that does not alter the DNA sequence). In this case, the methylation data come in the form of normalized methylation levels (M-values) where negative values correspond to unmethylated DNA and positive values correspond to methylated DNA. Along with this, we have a number of sample phenotypes (e.g., BMI, Sex, Age in years). This methylation object is a *"GenomicRatioset"*, a *Bioconductor* [22] data object derived from the *"SummarizedExperiment"* [20]. These *"SummarizedExperiment"* objects contain *"assays"*, in this case normalized methylation levels, and optional sample level *"ColData"* and feature-level *"metadata"*. These objects are very convenient to contain all of the information about a dataset in a high-throughput context. For more details on these objects, one could consult the *vignette on Bioconductor*.

After reading in the data, we can see in the provided R output that this object has $dim()$ of 5000×37, meaning it has 5000 features and 37 samples (observations). To extract the matrix of methylation M-values, we use the *"assay()"* function. Note that in the matrix of methylation data, samples or observations are stored as rows.

In this episode, we will focus on the association between **Age** and **methylation**.

b: Experiment steps:

Let's denote by X the methylation matrix,

- **1) Singularity:** We investigate the singularity of the matrix X^tX and check out what happens if we try to fit a linear model to the data.
- **2) Ordinary least square versus Ridge regression:** Here, we work with a set of features known to be associated with **Age** from a paper by *Horvath et al.*. Horvath et al. used methylation markers alone to predict the biological **Age** of an individual.
  - We extract the first 20 features identified by Horvath, investigate correlations, and split the methylation data matrix and the age vector into training and test sets.
  - We fit both linear regression and ridge regression on the training data matrix and training **Age** vector using the previous features and record the MSE between our predictions and the true **Age**s for the test data.
- **3) Apply regularization methods:** We perform the Lasso, Ridge, and Elastic-net on the whole DNA methylation data using cross-validation to select the tuning parameter, examine the coefficients paths for each method, and load Horvath signature to compare features selected by Lasso and the elastic-net methods.

c: Results.

1) We can see that we are able to get some effect size estimates, but they seem very high. The *"Summary"* also says that we were unable to estimate effect sizes for 4964 features because of singularities. What this means is that R couldn't find a way to perform the necessary calculations due to the fact that we have more features than observations.
2) Predictors are correlated with each other. Since we split the data into test and training data, we can see that ridge regression gives us a better prediction on unseen data despite being worse on train data: $MSE_{lm} = 45.14 \geq MSE_{ridge} = 25.30$.
3) Comparing the features selected by Lasso (41 features) and the elastic net (60 features) with Horvath signature, we can see that we selected some of the same features (8 features for Lasso and 11 features for elastic net).

V. Measurement Error In Regression Theory

In some sense, all statistical problems involve measurement error.

Measurement errors occur whenever we cannot exactly observe one or more of the variables that enter into a model of interest. There are many reasons such errors occur, the most common ones being "sampling error and instrument error." Where any notation is used here, the true value is denoted as *"X"* and the variable observed in place of *"X"* is denoted by *"W"* (error-prone measurement). When the true and observed values are both categorical, then measurement error is more specifically referred to as **misclassification**.

Measurement errors occur in nearly every discipline. Here is a collection of examples in the bio-medical field:

- Genomics: In recent decades, genetic and epigenetic studies have become increasingly more important in medical research, but the process of sequencing DNA typically involves some errors.
- Disease status: In epidemiology, the outcome variable is often the presence or absence of a disease such as breast cancer, hepatitis, AIDS, etc. This is often assessed through an imperfect diagnostic procedure such as an imaging technique or a blood test, which can lead to either false positives or false negatives (misclassification).

1) Objective and Some Terminology

- How to model measurement error?
- What are the effects of ignoring it?
- How, if at all, can we correct for measurement error?

These are three general objectives in the measurement error problem that we will try to address in this part of our work.

2) The Model Description

One of the fundamental assumptions in the linear regression analysis is that all observations are correctly observed. When this assumption is violated, measurement errors creep into the data. The usual statistical tools tend to lose their validity (*see [6] and [11] for more details*). An important issue in the area of measurement errors is to find consistent estimators of the parameters, which can be accomplished by using some additional information from outside the sample.

In sections *0.6* and *0.7*, we consider a linear regression model defined in (3) with additive error,

$$y = X\beta + \epsilon, \; W = X + U \tag{32}$$

$X_i = (X_{i1}, \ldots, X_{ip})^t, \quad W_i = (W_{i1}, \ldots, W_{ip})^t,$
$U_i = (U_{i1}, \ldots, U_{ip})^t;$

$$X = \begin{bmatrix} X_1^t \\ \vdots \\ X_n^t \end{bmatrix};\; U = \begin{bmatrix} U_1^t \\ \vdots \\ U_n^t \end{bmatrix};\; W = \begin{bmatrix} W_1^t \\ \vdots \\ W_n^t \end{bmatrix} \;\; (n \times p \; matrix)$$

For the sake of notation simplicity, we assume that $\beta_0 = 0$. The true covariates X are not observed, and instead we have noisy measurements $W = X + U$ where U is an $n \times p$ random noise matrix with covariance matrix Σ_U. If the k-th variable has been measured correctly, the corresponding column of U will be set equal to zero, as will the variance of the measurement error of the k-th variable, $\Sigma_{U(k,k)} = 0$.

Assumption

- The matrix of measurement error $U \in \mathbb{R}^{n\times p}$ is assumed to have normally distributed rows, with mean zero and covariance Σ_U.
- Furthermore, assume that ϵ and U are independent and Σ_U is a $p \times p$ matrix of known values with non-negative diagonal elements.

Remark V.1. *It follows from the structural model*

$$y_i = \beta^t X_i + \epsilon_i,\ W_i = X_i + U_i \tag{33}$$

that the vector $\left(y_i, W_i^t\right)^t$ *follows a* $p+1$*-variate normal distribution with mean* $\mu = \left(\beta^t\mu_X, \mu_X^t\right)^t$ *and the covariance matrix,*

$$\Gamma = \begin{bmatrix} \sigma_Y^2 & \Sigma_{YW} \\ \Sigma_{WY} & \Sigma_W \end{bmatrix} = \begin{bmatrix} \sigma^2 + \beta^t\Sigma_X\beta & \beta^t\Sigma_X \\ \Sigma_X\beta & \Sigma_X + \Sigma_U \end{bmatrix}. \tag{34}$$

This leads to:

$$y_i | W_i = w_i = \gamma^t w_i + \delta_i \tag{35}$$

where $\delta = (\delta_1, \ldots, \delta_n)^t$ are i.i.d normally distributed with mean zero and variance σ_δ^2.

Theorem V.1. *Under the given assumptions,* γ *and* σ_δ^2 *are given by,*

$$\gamma = \left(\Sigma_W\right)^{-1}\Sigma_X\beta = \left(\Sigma_X + \Sigma_U\right)^{-1}\Sigma_X\beta \tag{36}$$

$$\sigma_\delta^2 = \sigma^2 + \beta^t\Sigma_X\beta - \gamma^t\left(\Sigma_X + \Sigma_U\right)\gamma \tag{37}$$

Thus,

$$\beta = \mathcal{K}_X^{-1}\gamma. \tag{38}$$

where $\mathcal{K}_X = \left(\Sigma_X + \Sigma_U\right)^{-1}\Sigma_X$ is a $p \times p$ matrix referred to as the *reliability matrix, see Gleser (1992) [12] and Aickin and Ritenbaugh (1992)* for example, discussion and illustrations of the role of the reliability matrix.

3) Estimated Coefficients and Behavior of naive analyses

Statistical analysis that is carried out by ignoring the presence of the measurement error is called a naive approach.

Without measurement error, we saw that the estimated coefficients and the unbiased estimator of σ^2 are given by $\hat\beta = (X^tX)^{-1}X^ty$ (4) and
$\hat\sigma^2 = \frac{1}{n-p}\sum_i(y - \hat y_i)^2,\ with\ \hat y_i = \hat\beta^t x_i$.

Proposition V.1. *The maximum likelihood estimators of* γ *and* σ_δ^2 *are just the naive least squares estimators,*

$$\begin{aligned} &\hat\beta_{naive} = \hat\gamma = (W^tW)^{-1}W^ty = S_{WW}^{-1}S_{Wy} \\ &\hat\sigma^2_{naive} = \hat\sigma_{delta} = \frac{1}{n-p}\sum_i(y - \hat y_i)^2,\ with\ \hat y_i = \hat\beta^t_{naive}w_i \end{aligned} \tag{39}$$

where, $S_{WW} = \frac{W^tW}{n}$ is the unbiased estimator of Σ_W and $S_{Wy} = \frac{W^ty}{n}$

Proposition V.2. *The exact bias expression for the naive estimators under the given assumptions is given by:*

$$\mathbb{E}\left[\hat\beta_{naive}\right] = \gamma = \mathcal{K}_X\beta\ ,\ \mathbb{E}\left[\hat\sigma^2_{naive}\right] = \sigma_\delta^2 \tag{40}$$

Remark V.2. *This result leads to an important conclusion: The measurement error in one of the variables may induce bias in the estimation of all coefficients including those measured without error.If more covariates are affected by measurement error, the resulting bias may become rather complex and the effect of measurement error may become difficult to describe.*

4) Correcting for Measurement Error in Multilinear regression

With some exceptions (*see [2], chap11 and 12*), correcting for measurement error requires information or data as laid out in item **3)** section 1.

Myriad approaches to carrying out corrections for measurement error have emerged. A number of which are described in [2]. These include *direct bias correction, moment based approach, likelihood based techniques, SIMEX and techniques based on modifying equations..*

Proposition V.3. *When* Σ_U *is known and* $\mathcal{K}_X$ *is unknown, then* $\mathcal{K}_X$ *is estimated consistently by replacing* Σ_X *and* Σ_W *by their respective consistent estimators as:*

$$\begin{aligned} &\hat\Sigma_X = \hat\Sigma_W - \Sigma_U \\ &\hat\Sigma_W = S_{WW} = \frac{W^tW}{n};\quad and\ \hat{\mathcal{K}}_X = S_{WW}^{-1}(S_{WW} - \Sigma_U). \end{aligned} \tag{41}$$

Corollary 3. *The maximum likelihood estimates of* β *and* σ^2 *are given by :*

$$\begin{aligned} \hat\beta &= \hat{\mathcal{K}}_X^{-1}\hat\gamma = \left(S_{WW} - \Sigma_U\right)^{-1}S_{Wy} \\ \hat\sigma^2 &= \hat\sigma_\delta^2 - \hat\beta^t\Sigma_U\hat{\mathcal{K}}_X\hat\beta \end{aligned} \tag{42}$$

$\hat\beta$ is an unbiased estimator and its covariance is given by:

$$Cov(\hat\beta) = Cov(\hat{\mathcal{K}}_X^{-1}\hat\gamma) = \sigma_\delta\big(\underbrace{n\Sigma_X\Sigma_W^{-1}\Sigma_X}_{C}\big)^{-1} = \sigma_\delta C^{-1}$$

When measurement error is present and Σ_U is not known, it can be estimated through replicated measurements of W.

Proposition V.4. *Suppose on unit i there are $m_i > 1$ replicated values $W_{i1}, \dots, W_{im_i}$ of the error-prone measure of x and $\bar{W}_{i.} = \sum_{k=1}^{m_i} \frac{W_{ik}}{m_i}$ their mean. Replication allows us to estimate Σ_U as:*

$$\hat{\Sigma}_U = \frac{1}{n}\sum_{i=1}^{n} \frac{\sum_{k=1}^{m_i}\left(W_{ik} - \bar{W}_{i.}\right)\left(W_{ik} - \bar{W}_{i.}\right)^t}{m_i - 1} \tag{43}$$

In that case;

$$\begin{aligned} \hat{\Sigma}_X &= S_{WW} - \hat{\Sigma}_U,\ \hat{\mathcal{K}}_X = S_{WW}^{-1}\left(S_{WW} - \hat{\Sigma}_U\right) \\ \hat{\beta} &= \left(S_{WW} - \hat{\Sigma}_U\right)^{-1} S_{Wy} \end{aligned} \tag{44}$$

Remark V.3. *With sufficiently large measurement error, it is possible that $S_{WW} - \hat{\Sigma}_U$ can be negative. In that case, some adjustment must be made; see Block and Peterson (1975).*

Our discussion of the linear model is intended only to set the stage for our main topic, **measurement error in high-dimensional context**, and is far from complete. A vast literature exists on measurement error. There is a number of excellent books, starting with one by *Fuller [11]* who wrote the first influential book focusing on linear regression models, and one by *Caroll et al. [5]* who treated measurement error in a much broader application context. Another book that gives a wide treatment of the topic is by *Buanaccorsi [6]* who focuses on different topics from those in the aforementioned two books and places emphasis on a more applied approach.

VI. Measurement Error in High-Dimensional Context: Behavior and Correction Methods

A. Ridge Regression Estimation Over Measurement Error -Ridden Data.

The standard assumption in linear regression analysis is that explanatory variables are uncorrelated. When this assumption is violated, the explanatory variables are nearly dependent, which is referred to as the **multicollinearity problem** (very common in high-dimensional data) and yields poor estimators of interest parameters. In order to resolve this problem, several approaches have been considered. Among them, the "Ridge regression" introduced by *Horel and Kennard [15]* was discussed in section III and considers a shrinkage method to overcome the problem of multicollinearity for the estimation of regression parameters.

When the problem of multicollinearity is present in the measurement error-ridden data, an important issue is how to obtain consistent estimators of regression coefficients. One simple idea is to use ridge regression estimation over the error-ridden data. An obvious question that arises is: what happens then?

In this section, we will attempt to answer these questions.

1) Ridge Regression Estimator of β and its Asymptotic Properties.

Here we introduce the ridge regression estimators of β.For this, we first consider the conditional setup of the least squares method 32 with known *reliability matrix* $\mathcal{K}_X$.Remember in this case that the corrected moment estimator or corrected score estimated of β and γ are respectively given by:

$$\hat{\beta}^{LS}_{ME} = \mathcal{K}_X^{-1}\hat{\gamma} = \left(S_{WW} - \Sigma_U\right)^{-1} S_{Wy} \text{ (42)}$$

$$\textit{and } \gamma = \mathcal{K}_X \beta \text{ (36)}$$

Where "*ME*" stands for measurement error. The suggested estimator of β based on a shrinkage strategy is obtained by minimizing ,

$$\underset{\beta \in \mathbb{R}^p}{minimize}\left\{ \parallel y - W\gamma \parallel_2^2 \right\} \quad S.t \parallel \beta \parallel^2 \leq s \ (s\ constant) \tag{45}$$

the Lagrangian problem become

$$\underset{\beta \in \mathbb{R}^p}{minimize}\left\{ \parallel y - W\mathcal{K}_X\beta \parallel_2^2 + k \parallel \beta \parallel^2 \right\} \tag{46}$$

Proposition VI.1. *The numerical solution of this problem corresponding to the ridge regression estimator of β in measurement error model 32 is given by:*

$$\hat{\beta}^{R}_{ME} = \left[\mathbb{I}_p + k\left(n\mathcal{K}_X^t S_{WW}\mathcal{K}_X\right)^{-1}\right]^{-1} \hat{\beta}^{LS}_{ME}. \tag{47}$$

Corollary 4. *Substituting the consistent estimator of $\mathcal{K}_X$ given in* (41) *we get,*

$$\hat{\beta}^{R}_{ME} = \left[\mathbb{I}_p + k\left(n\hat{\mathcal{K}}_X^t S_{WW}\hat{\mathcal{K}}_X\right)^{-1}\right]^{-1} \hat{\beta}^{LS}_{ME}. \tag{48}$$

Denote the ridge factor of ridge estimation by: $Z_n^{ME} = \left[\mathbb{I}_p + kC_n^{-1}\right]^{-1}$ with $C_n = n\hat{\mathcal{K}}_X^t S_{WW}\hat{\mathcal{K}}_X$.

Corollary 5. *The mean square error of $\hat{\beta}^{R}_{ME}$ is given by:*

$$MSE(\hat{\beta}^{R}_{ME}, k) = k^2\beta^t\left[C_n + k\mathbb{I}_p\right]^{-2}\beta + \sigma_\delta^2 tr\left(Z_n^{ME} C_n^{-1}\left(Z_n^{ME}\right)^t\right) \tag{49}$$

Remark VI.1. *When $n \to \infty$*
then $C_n \to C$, $Z_n^{ME} \to Z^{ME}$ and

$$MSE(\hat{\beta}^{R}_{ME}, k) = k^2\beta^t\left[C + k\mathbb{I}_p\right]^{-2}\beta + \sigma_\delta^2 tr\left(Z^{ME} C^{-1}\left(Z^{ME}\right)^t\right)$$

if $k = 0$ then $Z^{ME} = \mathbb{I}_p$ and
$MSE(\hat{\beta}^{R}_{ME}, k) = \sigma_\delta^2 tr(C^{-1}) = MSE(\hat{\beta}^{LS}_{ME})$.

2) Comparison of $\hat{\beta}^{R}_{ME}$ and β^{LS}_{ME}

Let $\lambda_{max} = \lambda_1 \geq \cdots \geq \lambda_p = \lambda_{min} > 0$ denote the eigenvalues of the positive definite matrix $C = n\mathcal{K}_X^t\Sigma_W\mathcal{K}$.we can find and orthogonal matrix P such that, $P^tCP = D = diag(\lambda_1, \dots, \lambda_p)$ (see **Remark 1.2.1**); The corresponding eigenvalues of Z^{ME} and

$\left[C+k\mathbb{I}_p\right]^{-1}$ are respectively, $\frac{\lambda_j}{\lambda_j+k}$, $\frac{1}{\lambda_j+k}$ $j=1,\ldots,p$ so that.

$$\begin{aligned} k^2\beta^t\left[C_n+k\mathbb{I}_p\right]^{-2}\beta &= k^2\beta^tP^t\left[D+k\mathbb{I}_p\right]^{-2}P\beta \\ &= k^2\sum_{j=1}^{p}\frac{\alpha_j^2}{(\lambda_j+k)^2}, \\ & \quad \textit{where } \alpha = P\beta\ ,\ (p\times 1 \textit{ vector}) \end{aligned}$$

and

$$\sigma_\delta^2 tr\left(Z^{ME}C^{-1}(Z^{ME})^t\right) = \sigma_\delta^2\sum_{j=1}^{p}\frac{\lambda_j}{(\lambda_j+k)^2}$$

. Now the MSE of $\hat{\beta}^R_{ME}$ may be written as:

$$\begin{aligned} MSE(\hat{\beta}^R_{ME},k) &= k^2\sum_{j=1}^{p}\frac{\alpha_j^2}{(\lambda_j+k)^2}+\sigma_\delta^2\sum_{j=1}^{p}\frac{\lambda_j}{(\lambda_j+k)^2} \\ &= \psi_b(k)+\psi_v(k). \end{aligned} \tag{50}$$

Theorem VI.1 (*from [21]*)**.** *There always exist a $k>0$ such that ,*

$$MSE(\hat{\beta}^R_{ME},k) < MSE(\hat{\beta}^{LS}_{ME})\ . \tag{51}$$

VII. Measurement Error In Lasso

Modern statistics is facing problems due to the increase in dimensionality of the data in fields such as genomics, finance, network analysis,…etc. It is quite common in high-dimensional regression, where the number of variables p largely exceeds the sample size n, to assume that the number of covariates s that have an effect on the response variable y is much less than n (*sparsity assumption*). Hence, the vector of regression parameters is assumed to be s-sparse. A plethora of high-dimensional regression methods are available, among which are the Lasso regression [24], Dantzig selector (DS) [6], and Smoothly Clipped Absolute Deviation (SCAD) [11]. These methods all allow for model selection and parameter estimation through a penalization of the parameters, as seen in the case of Lasso. These methods are developed for the case in which the covariates are fully observed and without errors;. However, in many applications, our data are subject to at least some measurement error. In the classical regression context, when $p<n$ and standard methods can be applied, it is well known that measurement error in the covariates will lead to bias in the estimation of the parameters (40) and to loss of power [5]. Since the standard Lasso is widely used despite the presence of measurement error, it is of interest to study the effects measurement error has on the analysis and describe some of the statistical methods used to correct for those effects.

A. Impact Of Ignoring Measurement Error

The notation used to study the properties of lasso is used for W and U. We partition the variance matrix in the form:

$$S_{WW} = \begin{bmatrix} S_{WW}(S,S) & S_{WW}(S,S^c) \\ S_{WW}(S^c,S) & S_{WW}(S^c,S^c) \end{bmatrix} \tag{52}$$

We saw that in the absence of measurement error, the Lasso is consistent for prediction and estimation (24)). $y = X\beta+\epsilon = (W+U)\beta+\epsilon = W\beta+\underbrace{\epsilon-U\beta}_{\delta}.$

Proposition VII.1. *Assume the compatibility condition* (23) *holds with constant Φ , and that there exist a constant λ_0 such that $\frac{2}{n}\parallel\delta^tW\parallel_\infty\leq\lambda_0$; Then, with a regularization parameter $\lambda\geq 2\lambda_0$,*

$$\frac{1}{n}\parallel W(\hat{\beta}^{Lasso}-\beta)\parallel_2^2+\lambda\parallel\hat{\beta}^{Lasso}-\beta\parallel_1\leq\frac{4\lambda^2s}{\Phi_o^2}. \tag{53}$$

This shows that in the presence of measurement error, the estimation error of Lasso can be bounded.Using the triangle inequality, we have

$$\parallel\delta^tW\parallel_\infty\leq\parallel\epsilon^tW\parallel_\infty+\parallel\beta^tU^tX\parallel_\infty+\parallel\beta\parallel_1\parallel U^tU\parallel_\infty$$

Hence the bound (53) is implied by ,

$$\frac{2}{n}\parallel\epsilon^tW\parallel_\infty+\frac{2}{n}\parallel\beta^tU^tX\parallel_\infty+2\parallel\beta\parallel_1\parallel\frac{U^tU}{n}\parallel_\infty\leq\lambda_0\ ; \tag{54}$$

and the Lasso with measurement error is consistent if all the three terms in the above expression (54) converge to 0.However,

$$\frac{U^tU}{n}\underset{n\to+\infty}{\longrightarrow}\Sigma_U \quad and \quad \parallel\Sigma_U\parallel_\infty\neq 0$$

, Consequently, we do not achieve consistency.

We have just seen that standard results for consistency of estimation no longer hold when the covariates are affected by measurement error. Now let's see how measurement error affects covariate selection with Lasso. By definition (29), the "irrepresentable condition with measurement error."(**IC-ME**) hold if there exists a constant $\theta\in[0,1[$ such that ,

$$\parallel S_{WW}(S^c,S)S_{WW}(S,S)^{-1}sign(\beta_S)\parallel_\infty\leq\theta.. \tag{55}$$

In presence of measurement error, *Sorensen, Frigessi and Thoren (2015) [23]* shown that to achieve covariate selection consistency, we need the following additional condition called "Measurement Error Condition" (**MEC**):

Definition VII.1 (MEC)**.** *The measurement error condition (MEC) is satisfied if*

$$\Sigma_W(S^c,S)\Sigma_W(S,S)^{-1}\Sigma_U(S,S)-\Sigma_U(S^c,S)=0.\ ,\ (\textit{visit } [23]). \tag{56}$$

B. Correction for Measurement Error in Lasso

The objective of this section is to describe some penalized regression correction methods that can be employed to simultaneously correct both variable selection and model estimation. It is assumed that measurement error is adequately modeled, specifically "additive measurement error".

To demonstrate the bias in estimation resulting from measurement error, let's consider the naive Lasso approach. In this approach, we substitute W for X in the Lasso estimator defined in equation (18).

$$\hat{\beta}^{LS}(\lambda_n) = argmin_{\beta\in\mathbb{R}^p}\left\{ \parallel y - W\beta \parallel_2^2 + \lambda_n \parallel \beta \parallel_1 \right\}. \tag{57}$$

It is possible to demonstrate that this yield the bias loss function:

$$\mathbb{E}\left[\parallel y - W\beta \parallel_2^2 |X,y\right] = \parallel y - X\beta \parallel_2^2 + n\beta^t\Sigma_U\beta. \tag{58}$$

1) Corrected Lasso (Non Convex Lasso)

The must natural way for correcting for the bias in (58) leads to the constrained correct Lasso (CCL):

$$\hat{\beta}_{CCL} \in \underset{\beta:\|\beta\|_1\leq R}{argmin}\left\{\frac{1}{n} \parallel y - W\beta \parallel_2^2 - \beta^t\Sigma_U\beta\right\}. \tag{59}$$

or alternatively , the regularized version (regularize corrected Lasso),

$$\hat{\beta}_{RCL} \in \underset{\beta\in\mathbb{R}^p}{argmin}\left\{\frac{1}{n} \parallel y - W\beta \parallel_2^2 - \beta^t\Sigma_U\beta + \lambda_{RCL} \parallel \beta \parallel_1\right\}. \tag{60}$$

both introduced by *Loh and Wainright (2012) [23].*

Since in practice we may not know the covariance matrix Σ_X, given the set of samples, it is natural to form estimates of the quantities Σ_X and $\Sigma_X\beta$ as:

$$\hat{\Sigma_X} = \frac{W^tW}{n} - \Sigma_U \text{ , } and \text{ } \hat{\gamma} = \frac{1}{n}W^ty$$

. Let's note that in practice, Σ_U is unknown and needs to be estimated from data.

Proposition VII.2. *The estimator* (59) *and* (60) *can be reformulated as:*

$$\hat{\beta}_{CCL} \in \underset{\beta:\|\beta\|_1\leq R}{argmin}\left\{\frac{1}{2}\beta^t\hat{\Sigma}_X\beta - \hat{\gamma}^t\beta\right\}. \text{ , } and \tag{61}$$

$$\hat{\beta}_{RCL} \in \underset{\beta:\|\beta\|_1\leq b_0\sqrt{s}}{argmin}\left\{\frac{1}{2}\beta^t\hat{\Sigma}_X\beta - \hat{\gamma}^t\beta + \lambda_{RCL} \parallel \beta \parallel_1\right\}. \tag{62}$$

(b_0 *constant*)

Remark VII.1. *When* $\Sigma_U = 0_{\mathbb{R}^{p\times p}}$ *(corresponding to the noiseless case), the estimators reduce to the standard Lasso.However, when* $\Sigma_U \neq 0_{\mathbb{R}^{p\times p}}$*, the matrix* $\hat{\Sigma}_X$ *is not positive semi-definite in high-dimensional regime* ($p >> n$)*.Indeed, since the matrix* $\frac{1}{n}W^tW$ *has rank at must* n*, the subtracted matrix* Σ_U *may cause* $\hat{\Sigma}_X$ *to have a large number of negative eigenvalues.Consequently the quadratic losses appearing in the problems* (59) *and* (60) *are **non convex**.*

Remark VII.2. *When* $\hat{\Sigma}_X$ *has negative eigenvalues (which occur frequently in high-dimensional cases), the objective function in equation* (60) *is unbounded from below, hence we make use of the regularized estimator* (62) *to overcome these technical difficulties.*

Remark VII.3. *Note that,"*$\in$*" and not "*$=$*" has been used because in the presence of non-convexity, it is generally impossible to provide a polynomial-time algorithm that converges to a (near) global optimum due to the presence of local minima.*

*Loh and Wainwright [9] demonstrated that a simple "**project gradient descent algorithm**" applied to the problems* (59) *or* (62) *(if* b_0 *is properly chosen) converge with high probability to a small neighbourhood of the set of all global minimizers.*

Definition VII.2. *Project gradient descent is a standard way to solve constrained optimization problem.*

2) Convex Conditional Lasso

A clear drawback of the previous method is that it leads to a non-convex optimization problem.The ideal behind CoCoLasso is to intervene directly on $\hat{\Sigma}_X$, the estimated covariance matrix of X, with a transformation that will provide a "positive semi-definite" matrix.

We first introduce some necessary notations and model setup:

- For any square matrix $G = (g_{ij})_{i,j}$, we write $G > 0$ ($\geq$0) when it is positive (semi-) definite.
- Let $\parallel G \parallel_{\max} = \max_{i,j}|g_{ij}|$ denote the element-wise maximum norm.
- We assume that all variables are centred so that the intercept term is not included in the model.

We now define a nearest positive semi-definite matrix operator as follows:

For any square matrix G,

$$(G)_+ = \underset{G_1\geq 0}{argmin} \parallel G - G_1 \parallel_{\max} \tag{63}$$

This operator will project the matrix $\hat{\Sigma}_X$ into a space of semi-definite matrix selecting the nearest one. Then, by denoting $\tilde{\Sigma}_X = (\hat{\Sigma}_X)_+$, the convex conditional Lasso is define as:

$$\hat{\beta}_{CoCo} = \underset{\beta\in\mathbb{R}^p}{argmin}\left\{\frac{1}{2}\beta^t\tilde{\Sigma}_X\beta - \hat{\gamma}^t\beta + \lambda_{CoCo} \parallel \beta \parallel_1\right\} \tag{64}$$

Remark VII.4. *The matrix* $\tilde{\Sigma}_X$ *is always positive semi-definite by construction while* $\hat{\Sigma}_X$ *is guaranteed to be* *positive semi-definite only for $p < n$.Consequently, the optimization problem in* (64) *is guaranteed to be convex.*

Theorem VII.1 (Cholesky decomposition)**.** *Let A be a real-valued symmetric (semi-) positive-definite matrix; There exist a lower triangular matrix L with real and positive diagonal entries, such that,*

$$A = L^T L \tag{65}$$

Defining $\frac{1}{\sqrt{n}}\tilde{X}$ the Cholesky factor of $\tilde{\Sigma}_X$ (i.e $\frac{1}{n}\tilde{X}^t\tilde{X} = \tilde{\Sigma}_X$) and $\tilde{y}$ such that $\frac{1}{n}\tilde{X}^t\tilde{y} = \hat{\gamma} = \frac{1}{n}W^t y$, the estimator (64) can be reformulates as:

$$\hat{\beta}_{CoCo} = \underset{\beta\in\mathbb{R}^p}{argmin}\left\{\frac{1}{n} \parallel \tilde{y} - \tilde{X}\beta \parallel_2^2 + \lambda_{CoCo} \parallel \beta \parallel_1 \right\} \tag{66}$$

Remark VII.5. *This is a regular Lasso regression of $\tilde{y}$ and $\tilde{X}$ with penalization parameter λ_{CoCo} .It is of great advantage for the practical implementation.We can apply any standard Lasso algorithm as the coordinate descent algorithm [10] or Least angle regression [8] to obtain solution.*

3) Selecting The Tuning Parameter Under Measurement Error

The choose of the tuning parameter in penalized methods relies on *Cross-Validation*. In presence of measurement error, naive application of Cross-validation might lead to bias results.To elucidate, consider the usual K-folds Cross-validation for selecting optimal λ in the clean Lasso.

If we naively use the observed data (W, y) , then the cross-validated choice of λ is defined by minimizing ,

$$CV_{(K)} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_k} \parallel y_k - W_k\hat{\beta}_k(\lambda) \parallel_2^2 . \tag{67}$$

Even if we use CoCoLasso or NCL to compute $\hat{\beta}_k(\lambda)$ based on $W_{-k} and\ y_{-k}$, the above criterion is biased compared to (31) in the same way we shown that the loss function in (57) is a biased version of the one in (18).Observing that (31) is equivalent to:

$$\hat{\lambda} = \underset{\lambda}{argmin}\left\{\frac{1}{K}\sum_{k=1}^{K}\frac{1}{2}\hat{\beta}_k^t(\lambda)\Sigma_k\hat{\beta}_k(\lambda) - \gamma_k^t\hat{\beta}_k(\lambda)\right\}. \tag{68}$$

where $\Sigma_k = \frac{1}{n_k}X_k^t X_k$ and $\gamma_k = \frac{1}{n_k}X_k^t y_k$.

Since unbiased the unbiased surrogate $\hat{\Sigma}_k$ possibly has negative eigenvalues, using it will lead to a cross validation function unbounded from below.*Datta and Zou [7]* substituted Σ_k and γ_k with their projected and estimated counterparts $\tilde{\Sigma}_k = (\hat{\Sigma}_k)_+$ and $\hat{\gamma}_k$ .With this correction, the cross-validated λ is defined as:

$$\tilde{\lambda} = \underset{\lambda}{argmin}\left\{\frac{1}{K}\sum_{k=1}^{K}\frac{1}{2}\hat{\beta}_k^t(\lambda)\tilde{\Sigma}_k\hat{\beta}_k(\lambda) - \hat{\gamma}_k^t\hat{\beta}_k(\lambda)\right\}. \tag{69}$$

$\tilde{\lambda}$ is an unbiased estimator of λ.

VIII. Matrix uncertainty selector (MU-Selector)

So far, we have seen that corrected Lasso (NCL) (62) and CoCoLasso correct for measurement error by including the covariance of the measurement error Σ_U in the model, resulting in estimators with good theoretical properties. However, this quantity is assumed to be known, and in practice, it is usually unknown. Estimating the covariance matrix of the measurement error requires additional data, such as replicated measurements of the covariates, and can be computationally expensive or even unfeasible when the number of variables p increases.

An interesting alternative is the so-called *Matrix Uncertainty Selector* proposed by *Rosenbaum and Tsybakov [19]*.

We consider the model in (32).We typically assume that β is "s-sparse" where $1 \le s \le p$ is some integer. In what follows, we assume that ϵ and U satisfy the assumptions:

$$\frac{1}{n} \parallel W^t\epsilon \parallel_\infty \le \lambda \ \ and \ \ \parallel U \parallel_\infty \le \delta. \quad (\ with\ high\ probability\ ). \tag{70}$$

The "Matrix Uncertainty Selector" $\hat{\beta}_{MUS}$ is define as the solution of the minimization problem:

$$\begin{aligned}\min\Bigg\{ &\parallel \beta \parallel_1 : \beta \in \Theta, \frac{1}{n} \parallel W^t(y - W\beta) \parallel_\infty \\ &\le (1+\delta)\delta \parallel \beta \parallel_1 + \lambda \Bigg\},\end{aligned} \tag{71}$$

where $\Theta \subseteq \mathbb{R}^p$ is a given set characterizing the prior knowledge about β.

The problem (71) is a convex minimization problem and it reduces to linear programming if $\Theta = \mathbb{R}^p$ .Throughout this section, we assume for simplicity that all diagonal elements of the Gram matrix $\frac{1}{n}X^t X$ are equal to 1.

Proposition VIII.1 (solution existence)**.** *Under assumptions* (70) *, the feasible set of the convex problem* (71) *is non empty,*

$$\begin{aligned}&\Psi = \left\{\beta \in \Theta, \frac{1}{n} \parallel W^t(y - W\beta) \parallel_\infty \le (1+\delta)\delta \parallel \beta \parallel_1 + \lambda\right\} \\ &\Psi \neq \emptyset\end{aligned} \tag{72}$$

Remark VIII.1. *If $\delta = 0$ and $\Theta = \mathbb{R}^p$, the MU-Selector becomes the Dantzig selector* (30)*. The MU-Selector can be seen as an evolution of the Dantzig selector that can also take into account the measurement error in the model without needing any information about the measurement error variance, but rather by using a supplementary tuning parameter ("δ").*

IX. Numerical Study

A. Ridge under measurement error (simulation)

As discussed earlier, ridge regression (13) provides better estimators when facing the problem of multi-collinearity in our data. The purpose of this simulation is to evaluate the performance of the modified ridge estimation in (46) when the problem of multi-collinearity is present in the measurement error-ridden data. To this end, we will restrict ourselves particularly to the case where p¡n (low-dimensional data) with high correlations between covariates measured with error.

a: Simulation design:

We simulate data from the true model ,

$$y = X\beta + \epsilon \quad , \epsilon \rightsquigarrow \mathcal{N}(0,1) \ , \ p = 100 \ and \ n = 500$$

where X has been generated as $X \rightsquigarrow \mathcal{N}(0, \Sigma_X)$ with $\Sigma_X = (\rho_{ij})$ $(\rho_{ij} = 0.9^{|i-j|})$. All coefficients are set to 3, $\beta = (3, 3, \ldots, 3)^t$.The observed data were generated as ,

$$W = X + U, \quad where \ U \rightsquigarrow \mathcal{N}(0, \Sigma_U) \ with \ \Sigma_U = 0.75\mathbb{I}_p$$

. The simulated data was divided into a training and a test set. The four methods; *True OLS*[6] ($y \sim X$)(4), *corrected OLS*(42), *naive ridge*, and *modified ridge regression*(46) were used to fit a corresponding model to the training set. The fitted models were used to make predictions on the test set, and we computed the MSE and the PE (prediction error) on the test set. The procedure was repeated 100 times.

b: Simulation results:

We can see in table 2 that both the MSE and PE (on average) of the estimates $\hat{\beta}$ provided by the modified (corrected) ridge are lower than those of the other three methods. This means that the provided $\hat{\beta}$ is much more reliable considering the MSE (as mentioned in *theorem 3.1.2*) and PE. We also find out, in passing, that using the corrected version of OLS (42) in this setting (*"high-correlation with measurement error"*) would result in a pretty poor estimator given the MSE and PE.(table 2).

B. Under sparsity assumption with measurement error: NCL, CoCoLasso and MUS implementaiton.

The purpose of this section is to implement and investigate the performance of methods for measurement error correction previously presented on simulated high-dimensional regression problems with measurement error involve in the covariates. We will mainly investigate variables selection behavior and estimation error of the coefficients as $L_2 - error,\ (mse) L_1 - error(l_1)\ and\ prediction\ error$. The R software has been used for the task.

1) Process/Simulation design:

Simulate data from the true model ,

$$y = X\beta + \epsilon \quad , \epsilon \rightsquigarrow \mathcal{N}(0, \frac{0.05}{1.96}),$$
$$p = 5000 \ and \ n = 200 \qquad (p >> n)$$

where X has been generated as $X \rightsquigarrow \mathcal{N}(0, \Sigma_X)$ with $\Sigma_X = (\rho_{ij})$ $(\rho_{ij} = 0.5^{|i-j|})$.The active set index have been chosen randomly and all the coefficients belonging to the active set S of dimension s are set to 2 (i.e $\beta_i = 2 \ if \ i \in S \ and \ 0 \ otherwise$) .The observed data were generated as ,

$$W = X + U, \quad where \ U \rightsquigarrow \mathcal{N}(0, \Sigma_U) \ with \ \Sigma_U = 0.75\mathbb{I}_p$$

.

For two differents size of $s \in \{5, 10\}$ (number of non-zero coefficients),for each simulation (100 simulations), we

- Train the naive Lasso model (on error prone data W) and the true Lasso model (on X) in addition to the three correction methods. [17]
- record the overall number of variables selected (nVS), the number of correct variables selected ($nCVS$, variables associated with non-zero coefficients), the mean square error (MSE), the $L_1 - error$ error as well as the prediction error (PE) of the estimate $\hat{\beta}$.

The results are summarized as average values over the number of simulation.

2) Implementation

The Lasso algorithm has been implemented using the *glmnet R package* choosing the optimal parameter λ based on cross-validation over 100 differnt values of λ in the interval $[10^{-2}, 10^2]$.

The corrected Lasso (NCL), Dantzig Selector and MUS have been implemented using the *hdme R package* [22]. In addition, the hyperparameter R (radius) in NCL has been selected based on Cross-validation procedure in the interval [$R_{max}/500; 2 \parallel \beta_n\hat{aive} \parallel_1$] as suggested by *Sorensen et al.* and *Datta and Zou* [7]; Regarding MUS, we selected λ as the Cross-validate estimated from naive Lasso $\lambda_n\hat{aive}$ and we choose δ according to the $Elbowrule$.

Finally, we implement CoCoLasso using *BDCoCoLasso R package* by Celiaescribe.The three correction methods have been performed with known covariance error matrix Σ_U .

3) Results:

For s=5 (Table. 3 and figure.2 & 3)

When considering estimation error (MSE, L_1-error, PE), the NCL and CoCoLasso tend to outperform the naive Lasso. In some cases, the difference is particularly significant with NCL. However, the MUS appears to be ineffective for estimation, resulting in higher metric values (MSE, PE) compared to the naive Lasso.

In terms of variable selection, while both the naive Lasso and the three correction methods generally

[6] Ordinary least square

succeed in identifying all non-zero coefficients, the NCL and MUS excel in accurately detecting all coefficients of the active set S with minimal false positives. In contrast, the naive Lasso tends to select too many false positives. On the other hand, the CoCoLasso tends to select an excessive number of variables, surpassing even the naive Lasso.

For s=10 (Table. 4 and figure.4 & 5)

Both NCL and MUS perform worse than the naive Lasso in terms of both estimation errors (MSE, L_1-error) and variable selection. They tend to choose too few variables, sometimes even fewer than the dimension of the active set S, resulting in false positives. However, in this specific situation, CoCoLasso generally outperforms the naive Lasso, NCL, and MUS in terms of both estimation error and model selection.

4) Discussion:

The results of this study are specific to a configuration where Σ_U is known and diagonal, meaning there is no correlation among measurement errors. Therefore, these findings cannot be applied universally, as different configurations may lead to different results.

However, based on our findings, NCL and MUS work well for model selection when the dimension of the active set is small ($s = 5$), and may be the best choice from a variable selection perspective. When the size of the active set increases ($s = 10$), the CoCoLasso model may have better performance than NCL and MUS. Thus, there is no globally preferable correction method to the naive Lasso estimator. In our specific case, the results vary depending on the size of the active set, and may change even more when exploring different measurement error structures such as correlated measurement error, estimated unknown $\hat{\Sigma_U}$, non-Gaussian measurement error assumption, etc.

APPENDIX

R codes, notebooks, graphics as well as data used for this work can be found here (github).

REFERENCES

[1] P. Bühlmann and S. Van De Geer. Statistics for high-dimensional data: methods, theory and applications. Springer Science & Business Media, 2011.

[2] J. P. Buonaccorsi. Measurement error: models, methods, and applications. CRC press, 2010.

[3] E. Candes and T. Tao. The dantzig selector: Statistical estimation when p is much larger than n. 2007.

[4] R. J. Carroll, D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu. Measurement error in nonlinear models: a modern perspective. Chapman and Hall/CRC, 2006.

[5] R. J. Carroll, D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu. Measurement error in nonlinear models: a modern perspective. Chapman and Hall/CRC, 2006.

[6] C.-L. Cheng and J. W. Van Ness. Statistical regression with measurement error. (No Title), 1999.

[7] A. Datta and H. Zou. Cocolasso for high-dimensional error-in-variables regression. 2017.

[8] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. 2004.

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

[10] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. Journal of statistical software, 33(1):1, 2010.

[11] W. A. Fuller. Measurement error models. John Wiley & Sons, 2009.

[12] L. J. Gleser. The importance of assessing measurement reliability in multivariate regression. Journal of the American Statistical Association, 87(419):696–707, 1992.

[13] T. Hastie, J. Qian, and K. Tay. An introduction to glmnet. CRAN R Repositary, 5:1–35, 2021.

[14] T. Hastie, R. Tibshirani, and M. Wainwright. Statistical learning with sparsity: the lasso and generalizations. CRC press, 2015.

[15] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1):55–67, 1970.

[16] J. Huang and C. X. Ling. Using auc and accuracy in evaluating learning algorithms. IEEE Transactions on knowledge and Data Engineering, 17(3):299–310, 2005.

[17] J. Luo, L. Yue, and G. Li. Overview of high-dimensional measurement error regression models. Mathematics, 11(14):3202, 2023.

[18] R. H. Myers and D. C. Montgomery. A tutorial on generalized linear models. Journal of Quality Technology, 29(3):274–291, 1997.

[19] M. Rosenbaum and A. B. Tsybakov. Sparse recovery under matrix uncertainty. The Annals of Statistics, pages 2620–2651, 2010.

[20] K. Rue-Albrecht, F. Marini, C. Soneson, and A. T. Lun. isee: interactive summarizedexperiment explorer. F1000Research, 7, 2018.

[21] A. M. E. Saleh et al. A ridge regression estimation approach to the measurement error model. Journal of Multivariate Analysis, 123:68–84, 2014.

[22] O. Sorensen. The hdme package: regression methods for high-dimensional data with measurement error.

[23] Ø. Sørensen, A. Frigessi, and M. Thoresen. Measurement error in lasso: Impact and likelihood bias correction. Statistica sinica, pages 809–829, 2015.

[24] R. Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1):267–288, 1996.

[25] X. Yan and X. Su. Linear regression analysis: theory and computing. world scientific, 2009.

[26] P. Zhao and B. Yu. On model selection consistency of lasso. The Journal of Machine Learning Research, 7:2541–2563, 2006.

models	**Example 1**			**Example 2**		
	AUC	ME	Nb. of $\hat{\beta} \neq 0$	AUC	ME	Nb. of $\hat{\beta} \neq 0$
Ridge	0.76 (0.042)	0.32 (0.053)	1000	0.76 (0.050)	0.31 (0.052)	1000
Lasso	0.65 (0.106)	0.41 (0.094)	29	0.55 (0.058)	0.46 (0.056)	20
Elastic Net	0.75 (0.062)	0.37 (0.074)	316	0.70 (0.076)	0.37 (0.068)	329
models	**Example 3**			**Example 4**		
	AUC	ME	Nb. of $\hat{\beta} \neq 0$	AUC	ME	Nb. of $\hat{\beta} \neq 0$
Ridge	0.92 (0.028)	0.16 (0.041)	1000	0.76 (0.047)	0.31 (0.041)	1000
Lasso	0.84 (0.037)	0.24 (0.048)	54	0.58 (0.073)	0.46 (0.070)	21
Elastic Net	0.90 (0.033)	0.17 (0.045)	415	0.70 (0.059)	0.36 (0.054)	361

TABLE 1. Simulation results.*The table reports the AUC, ME-values and number of non-zero* $\hat{\beta} - coefficients$ *.The simulation was repeated 100 times for each example and all results are reported as median values and (standard deviation sd.)*

	true OLS	**corrected OLS**	**naive Ridge**	**corrected Ridge**
MSE	6.85 (0.328)	477.94 (4700)	6.84 (0.327)	6.83 (0.327)
PE	2.01(0.146)	50058. (499396)	0.05 (0.008)	0.04 (0.006)

TABLE 2. *Simulation results for ridge under measurement error. The table reports the PE and the estimation error as* l_2 *norm (MSE).results are reported as median values and (standard deviation sd.)*

	True lasso	Naive lasso	True Dantzig	NCL	CoCoLasso	MUS
mse	6.10^{-6} (2.10^{-6})	0.126 (0.025)	6.10^{-6} (3.10^{-6})	0.104(0.025)	0.120 (0.026)	0.141 (0.018)
L_1-error	0.055 (0.011)	10.698 (2.3)	0.056 (0.012)	7.409 (2.67)	11.659 (1.90)	9.114 (0.881)
pe	0.0005 (1.10^{-3})	8.867 (3.4)	0.0005 (1.10^{-3})	8.080 (4.28)	7.121 (3.05)	12.599 (2.19)
nVS	8	34	28	9	45	15
$nCVS$	5	5	5	5	5	5

TABLE 3. *results after 100 simulations for s=5 (non-zero coefficients); p=5000 and n=200 reported as mean values and (standard deviation)*

	True lasso	Naive lasso	True Dantzig	NCL	CoCoLasso	MUS
mse	25.10^{-4} (2.10^{-6})	0.343 (0.04)	35.10^{-4} (2.10^{-5})	0.349(0.05)	0.333 (0.051)	0.353 (0.03)
L_1-error	0.167 (0.01)	21.892 (3.18)	0.207 (0.08)	19.602 (2.35)	22.64 (2.79)	20.211(1.39)
pe	0.001 (13.10^{-3})	24.343 (7.28)	0.001 (3.10^{-3})	26.902 (5.97)	21.114 (7.17)	29.164 (4.47)
nVS	25	29	41	8	38	18
nCV	10	5	10	3	7	5

TABLE 4. *results after 100 simulations for s=10 (non-zero coefficients); p=5000 and n=200 reported as mean values and (standard deviation)*

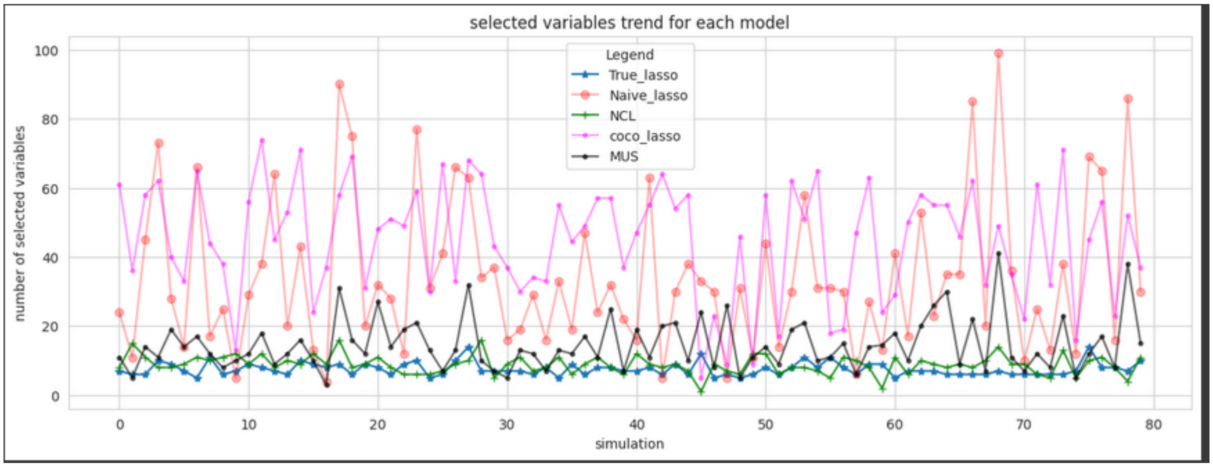


FIGURE 2. For s=5: The variable selection behavior paths of each models observed over 80 simulations.

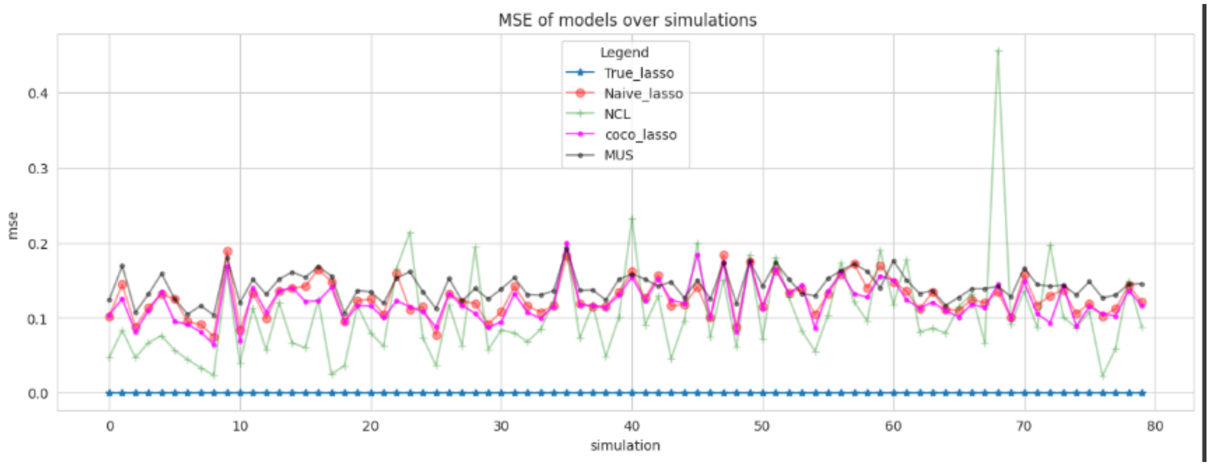


FIGURE 3. **For s=5: The Mean square error paths of models observed over 80 simulations.**

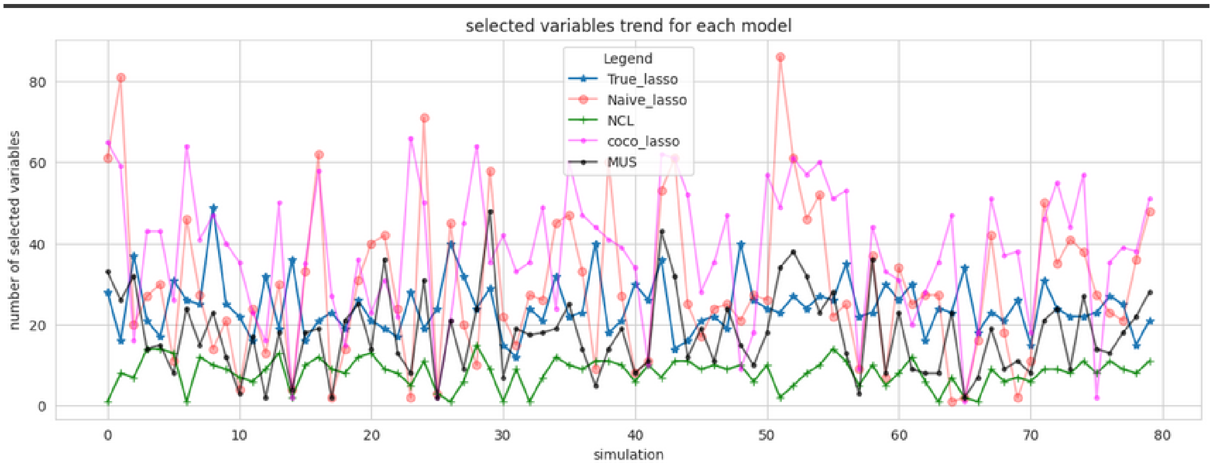


FIGURE 4. **For s=10: The variable selection behavior paths of models observed over 80 simulations.**

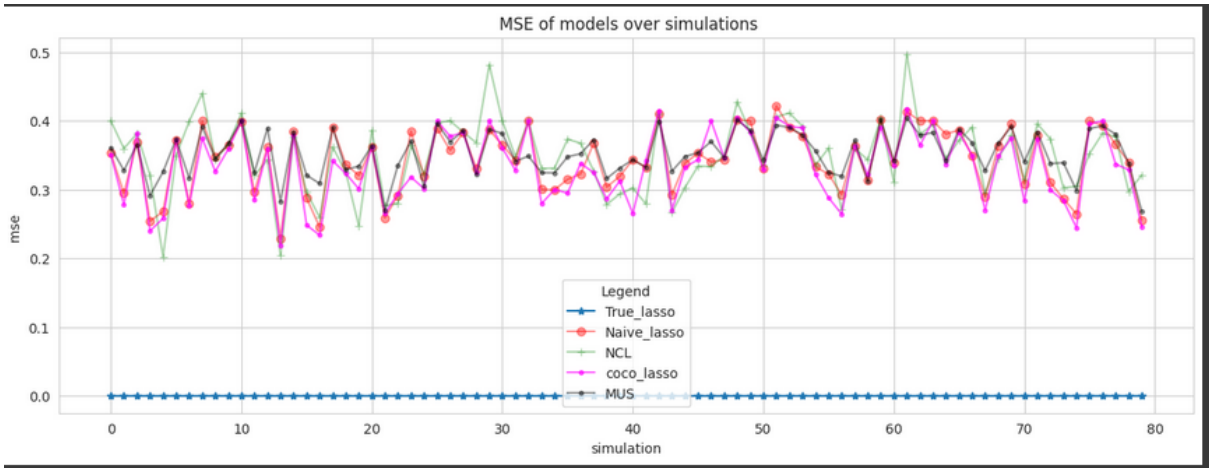


FIGURE 5. **For s=10: The Mean Square error paths of models observed over 80 simulations.**