Measurement Error In High-Dimensional Data

Abstract

In many important statistical applications, the number of variables (covariates) p largely exceed the number of observations (sample size) n .Such data are refer as high-dimensional data; preceding studies show that some common standard statistical methods of analysis do not conform with this kind of data.Besides, much applied work has been devoted to high-dimensional regression with clean data.However, we often face corrupted data in many applications like in genomic for instance where measurement error cannot be ignored.It is therefore necessarily to highlight some appropriate statistical methods for handling high-dimensional data as well as their extension to the case in which covariates are also mismeasured. The purpose of this study is to introduce reduction of high dimensional data using regularization methods (penalized linear regression) as well as their variants to accommodate for presence of measurement errors in covariates. In this paper, we evaluate four penalization methods ; ridge regression, Lasso, Dantzig selector and the Elastic net for model fitting then we present their respective variants to account for measurement error.The evaluation focus on situation relevant for practical applications through both simulated and real examples of high-dimensional data sets.

Keywords

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

0.1 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 had a tremendous growth in popularity and a plethora of methods has been proposed for statistical modelling of, and inference in high-dimensional data.Penalized regression methods such as ridge regression [18], Lasso [23] methods and Dantzig selector [5] are particularly good in this context.

In almost all disciplines, it may not be possible to observe a variable accurately, for some reason, and 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 the non-existent to the 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 on 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 creep into the data, there are tree main reason why measurement error cannot be ignored; it can cause bias in parameter estimation [4], interfere with variable selection [22] and lead to a loss of power [6] leading to 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 result in faulty inference as demonstrated for the Lasso [21]. Consequently, correction for measurement error in penalized regression has recently been studied by various authors.Example include ; "Ridge regression approach to measurement error"[21] , Non Convex Lasso (NCL) by Loh and Wainwright [22], the Convex Conditional Lasso (CoCoLasso) of Datta and Zou [9] and the Matrix Uncertainty Selector proposed by Rosenbaum and Tsybakov (MUS) [20].

The organization of this paper is as follows; **section 0.2** and **0.3** presents high-dimensional data together with potential challenges when analysing the later, along with some statistical methods one may use to handle this kind of datasets. **section 0.5** introduces the measurement error in regression theory, provides an overview of the consequences of measurement error in linear regression and introduces some corrections methods. **section 0.6** describes behaviour of measurement error in high-dimensional regression and introduces some high-dimensional approaches (methods) to correct for measurement error in high-dimensional context. Both real and simulated data are used for illustrations.

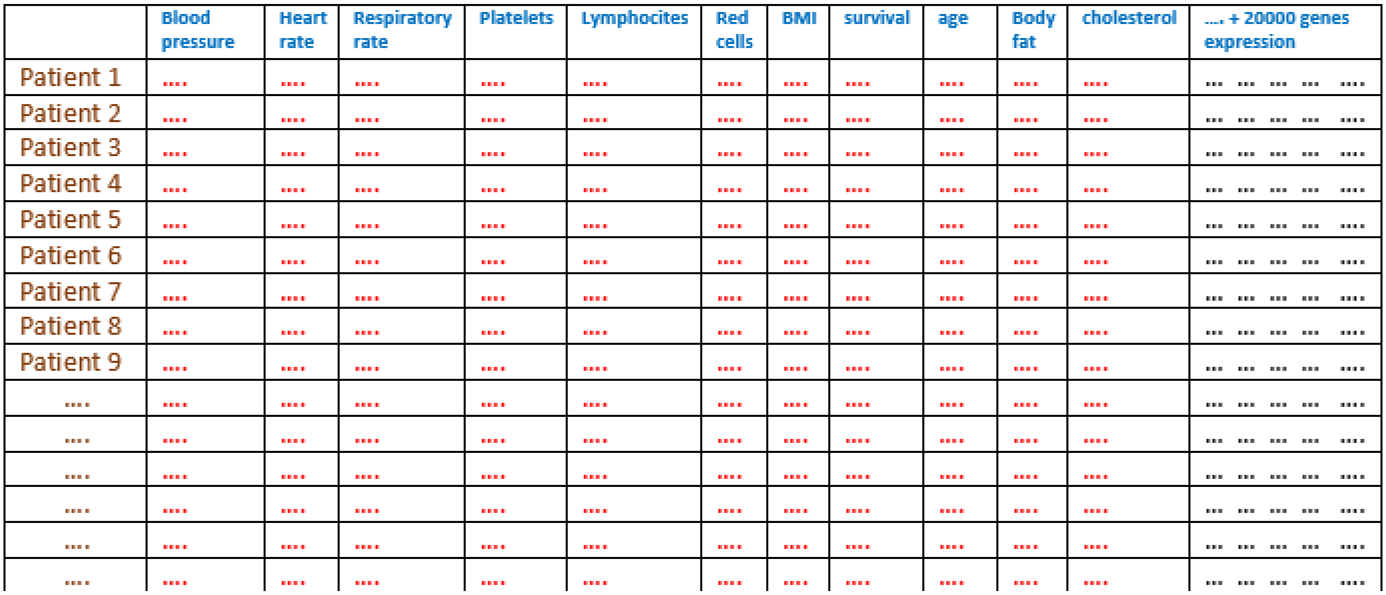
0.2 Introduction to High-Dimensional Data

High-dimensional data are defined as data in which the number of features (*variables observed*) **p**, are close to or large 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 feature **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 large number of observations. This concept should not be therefore confuse with notion of **big data** which is data that contains greater *variety*, arriving in increasing *volumes* and with more *velocity* known as the threes **Vs** (visit, `https://www.oracle.com/big-data/what-is-big-data/`).

High-dimensional datasets are become more common in many scientific fields as new automated data collection techniques have been developed.And example in biological sciences may include *data collected from hospital patients recording symptoms, blood test results, behaviours and general health* resulting in datasets with large number of features.
And 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 associate datasets can be considered as high-dimensional data:



	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

- predicting patient blood pressure using: *cholesterol level in blood,age and BMI as well as information on 200000 single nucleotide polymorphisms from 100 patients*

- Predicting probability of a patient's cancer progressing using: *gene expression data from 20000 genes as well as data associated with general patient health (age, weight,BMI, blood pressure) and cancer growth (tumour , localised spread,blood test results)*

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

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

Setup of Linear Regression Model

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

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

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

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

Where X_{ij} is the j^{th} variable for individual i and $\epsilon_i^{\prime} s$ 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:

$$y = X\beta + \epsilon, \tag{3}$$

where:

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \ X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \ \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}, \ and \ \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

The regression parameter are estimated by minimizing ordinary least squares:

$$\sum_{i=1}^{n}[y_i - (\beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip})]^2 = (y - X\beta)^t(y - X\beta) = \parallel y - X\beta \parallel^2, \ (\parallel . \parallel ^1).$$

Ordinary Least Squares Estimates (OLS Estimates)

Proposition 0.2.1 (*from [24]*). *The least squares estimation of β for linear regression model is given by,*

$$b = argmin_\beta \Big\{ \parallel y - X\beta \parallel_2^2 \Big\} = (X^tX)^{-1}X^ty, \tag{4}$$

assuming (X^tX) is a non-singular matrix.Note that this is equivalent to assuming that the matrix X is of full rank[2].
The estimator $b = (X^tX)^{-1}X^ty$ is and unbiased estimator of β. In addition,its covariance matrix is given by,

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

Proposition 0.2.2 (*from [24]*). *The unbiased estimator of the variance σ^2 in the multiple linear regression is given by:*

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

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:

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

[1] $\parallel . \parallel$ is the Euclidian norm on $\mathbb{R}^n$

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

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

Theoretically: When $p > n$, X^tX 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 0.2.3 (*from [24]*). *The average Euclidean distance measure $\mathbb{E}[\| b - \beta \|^2]$ between the least squares estimate b and the true parameter β is given by:*

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

Remark 0.2.1. *Assuming that (X^tX) has k distinct eigenvalues $\lambda_1, ..., \lambda_k$, then the eigenvalues of $(X^tX)^{-1}$ are $\frac{1}{\lambda_1}, ..., \frac{1}{\lambda_k}$, denoting by $V = (v_1, ..., v_k)^t$ the corresponding normalized eigenvectors, we can write $V^t(X^tX)^{-1}V = D = diag(\frac{1}{\lambda_1}, ..., \frac{1}{\lambda_k})$.*
Moreover, $tr(X^tX)^{-1}) = tr(V^tV(X^tX)^{-1}) = tr(V^t(X^tX)^{-1}V) = tr(D) = \sum_{i=1}^{k} \frac{1}{\lambda_i}$; we then have:

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

Now it is easy to see that if one of λ_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 column in X then, from *lemma(4)* , the covariance matrix $Cov(b) = (X^tX)^{-1}\sigma^2$ will have one or more large eigenvalues so that the mean Euclidean distance of $\mathbb{E}[\| b - \beta \|^2]$ will be inflated.Consequently, this makes the estimation of regression parameter β less reliable.Thus the high levels correlation between variable in high-dimensional datasets will have negative impact on least square estimates of regression parameter.

Clearly, alternative approaches that are better-suited to the high-dimensional setting are required.

0.3 Some Statistical Suitable Methods for Handling High-Dimensional Data

0.3.1 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 correlate 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 0.3.1. $\mathbb{E}[\parallel b - \beta \parallel^2] = \sum_{j=1}^n (\mathbb{E}[b_j] - \beta_j)^2 + \sum_{j=1}^n Var[b_j]$

According to "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 type of estimators, let $\hat{\beta}^{LS}$ denote the ordinary least square estimator of β. We shown that $MSE(\hat{\beta}^{LS}) = \mathbb{E}[\parallel \hat{\beta}^{LS} - \beta \parallel^2] = \sigma^2 tr[(X^tX)^{-1}]$ (8) thus, $\mathbb{E}[\parallel \hat{\beta}^{LS} \parallel^2] = \parallel \beta \parallel^2 + \sigma^2 tr[(X^tX)^{-1}$ (9); it can be seen that, with ill-conditioned X^tX, the resultant LSE $\hat{\beta}^{LS}$ would be large in length $\parallel \hat{\beta}^{LS} \parallel$ 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 biased estimator yet with small variance.

Theorem 0.3.1. *For any estimator b, the least squares criterion $\mathcal{Q}(b) = \parallel y - Xb \parallel^2$ can be rewritten as its minimum, reached at $\hat{\beta}^{LS}$ plus a quadratic form in b.*

$$\mathcal{Q}(b) = \parallel y - Xb \parallel^2 = \underbrace{\parallel y - X\hat{\beta}^{LS} \parallel^2}_{\mathcal{Q}_{min}} + \underbrace{(\hat{\beta}^{LS} - b)^t X^t X (\hat{\beta}^{LS} - b)}_{\phi(b)} = \mathcal{Q}_{min} + \phi(b) \tag{10}$$

contour for each constant of the quadratic form $\phi(b)$ are hyper-ellipsoids centred at ordinary LSE $\hat{\beta}^{LS}$. The optimization problem in Ridge regression can be state as:

$$minimize \parallel \beta \parallel^2 \ \ subject\ to\ (\hat{\beta}^{LS} - b)^t X^t X (\hat{\beta}^{LS} - b) = \phi_0\ for\ some\ constant\ \phi_0.$$

The enforced constrain 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

$$minimizing\ f(\beta) = \parallel \beta \parallel^2 + \frac{1}{k}[(\hat{\beta}^{LS} - b)^t X^t X (\hat{\beta}^{LS} - b) - \phi_0], \qquad k > 0$$

Where $\frac{1}{k}$ is the multiplier chosen to satisfy the constraint.

Proposition 0.3.2 (Hoerl and Kennard (1970))**.** *The numerical solution of this problem corresponding to the Ridge regression estimator of β is,*

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

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

$$minimize \ \parallel y - X\beta \parallel^2 \quad subject\ to\ \parallel \beta \parallel^2 \leq s\ for\ some\ constant\ s \tag{12}$$

the Lagrangian problem become

$$minimizing \ \parallel y - X\beta \parallel^2 + \lambda \parallel \beta \parallel^2 \tag{13}$$

which yield the same estimator given in (11).The penality parameter $\lambda \geq 0$ controls the amount of shrinkage in $\parallel \beta \parallel^2$.The large value of λ, the greater amount of shrinkage.For this reason, the Ridge estimator is also called the shrinkage estimator. There is one-to-one correspondence among $\lambda,\ s,\ k\ and\ \phi_0$.
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 0.3.3. *If $MSE(\hat{\beta}^R, k)$ denote the mean square error of ridge regression estimator, then*

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

Theorem 0.3.2 (*Hoerl and Kennard (1970)*)**.** *There always exists a $k > 0$ such that,*

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

0.3.2 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 $\lambda \parallel \beta \parallel^2$ 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 setting 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 disad-

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

vantage .The Lasso estimator of β is obtained by

$$minimizing \left\{ \parallel y - X\beta \parallel_2^2 \right\} \quad subject\ to \sum_{j=1}^{p} |\beta_j| \leq s\ for\ some\ constant\ s \tag{15}$$

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 become:

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

Computation of Lasso Solution

The Lasso problem is a convex program, specifically a quadratic program (**QP**) (*visit [17] 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 an effective computational algorithm, that gives insight into how the Lasso works. The Lagrangian form (16)is especially convenient for numerical computation of the solution.

Theoretical properties of Lasso penalty

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

Let $S = \{j : \beta_j \neq 0\}$ the index set of non-zero components of the true coefficient vector $\beta \in \mathbb{R}$ and denote the number of relevant covariate by $s = card\{S\}$.Under 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, $\hat{S}(\lambda) = \{j : \hat{\beta}_j(\lambda) \neq 0\}$.Given β ,we order the covariates such that $S = \{1, \ldots, s\}$, $S^c = \{s+1, \ldots, p\}$ and considering the partitioning $X = \left(X_S, X_{S^c}\right)$ 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 matrix are denote by Σ_X and the empirical covariance is given by $S_{XX} = \frac{X^T X}{n}$.

State the following basic inequality,

Lemma 5 ([3],P.103). *we have,*

$$\frac{1}{n} \parallel X\hat{\beta}^{Lasso} - X\beta \parallel_2^2 + \lambda \parallel \hat{\beta}^{Lasso} \parallel_1 \leq \frac{2\epsilon^t X(\hat{\beta}^{Lasso} - \beta)}{n} + \lambda \parallel \beta \parallel_1 . \tag{17}$$

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

$$\mathbb{P}(\mathcal{A}) \geq 1 - 2\exp(-\frac{t^2}{2}) \tag{18}$$

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:*

$$\frac{2}{n} \parallel X(\hat{\beta}^{Lasso} - \beta) \parallel_2^2 \leq 3\lambda \parallel \beta \parallel_1 \tag{19}$$

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

$$\beta_{j,S} = \beta_j \mathbb{1}_{\{j\in S\}}, \quad \beta_{j,S^c} = \beta_j \mathbb{1}_{\{j\notin S\}}. \tag{20}$$

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 0.3.1 (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*

$$\parallel \beta_S \parallel_1^2 \leq \frac{1}{n}\frac{s \parallel X\beta \parallel_2^2}{\Phi_o 2^2} = \frac{s(\beta^t S_{XX}\beta)}{\Phi_o^2}. \tag{21}$$

Theorem 0.3.3 (*[3], Theorem 6.1,P.107*)**.** *Suppose the compatibility condition holds for S.Then on $\mathcal{A}$, we have for $\lambda \geq 2\lambda_0$*

$$\frac{1}{n} \parallel X(\hat{\beta}^{Lasso} - \beta) \parallel_2^2 + \lambda \parallel \hat{\beta}^{Lasso} - \beta \parallel_1 \leq \frac{4\lambda^2 s}{\Phi_o^2}. \tag{22}$$

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

$$\frac{2}{n} \parallel X(\hat{\beta}^{Lasso} - \beta) \parallel_2^2 + \lambda \parallel \hat{\beta}_{S^c}^{Lasso} \parallel_1 \leq 3\lambda \parallel \hat{\beta}_S^{Lasso} - \beta_S \parallel_1 . \tag{23}$$

Remark 0.3.1. *The theorem combines two results:*

$$\frac{2}{n} \parallel X(\hat{\beta}^{Lasso} - \beta) \parallel_2^2 \leq \frac{4\lambda^2 s}{\Phi_o^2}, \quad (the\ bound\ for\ predictions\ error) \tag{24}$$

$$\parallel \hat{\beta}^{Lasso} - \beta \parallel_1 \leq \frac{4\lambda^2 s}{\Phi_o^2}, \quad (the\ bound\ for\ L_1 - error\ of\ coefficients\ estimates.) \tag{25}$$

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 $\lambda \approx \sqrt{\frac{log(p)}{n}}$,*

$$\parallel \hat{\beta}^{Lasso} - \beta \parallel_1 \xrightarrow[n \to +\infty]{\mathbb{P}} 0\,; \quad \parallel \hat{\beta}^{Lasso} - \beta \parallel_2 \xrightarrow[n \to +\infty]{\mathbb{P}} 0 \tag{26}$$

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) [25]* 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 $\mathbb{P}(\{\hat{S} = S\}) \xrightarrow[n \to +\infty]{\mathbb{P}} 1$. 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*[4] *consistency"* that does not assume the estimates to be estimation consistent.

Definition 0.3.2. *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 0.3.3. *Lasso is strongly sign consistent if there exists $\lambda_n = f(n)$, that is , a function independent of Y and X such that:*

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

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

Definition 0.3.4. *Lasso is general sign consistentcy if*

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

Remark 0.3.2.

- *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.*
- $(*) \Rightarrow (**)$

Definition 0.3.5 (Irrepresentable Condition)**.** *We say that , Irrepresentable condition is met for the set S if there exists a constant $\theta \in [0, 1[$ such that ,*

$$\| \ S_{XX}(S^c, S)S_{XX}(S, S)^{-1}sign(\beta_S) \ \|_\infty \leq \theta.. \tag{27}$$

Theorem 0.3.4 (Variables selection consistency,[*Zao and Yu (2006)*)**.** *] The irrepresentable condition* (27) *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) [25]* or *Meinshausen and Buhlmann (2010)* for more details.

Remark 0.3.3. *The irrepresentable condition , as given in* (27) *depends on the Gram matrix $\frac{X^tX}{n}$ but also on the signs of the true unknown parameter β, whereas the compatibility condition* (21) *only depends on Σ_X.*

0.3.3 Dantzig Selector (DS)

The Lasso is not the only $L_1 - penalization$ possible. from the score equation ,the Dantzig Selector by *Candes and Tao [5]* also belongs to the class of regularisation 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:

$$\underset{\beta\in\mathbb{R}^p}{min}\{ \ \| \ \beta \ \|_1 \ \} \ subject \ to \ \ \| \ X^t\big(y - X\beta\big) \ \|_\infty := \sup_{1\leq i \geq p} |\big(X^tr\big)_i| \leq \lambda_p.\sigma, \tag{28}$$

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

Remark 0.3.4. *The constraint on the residual vector imposes that for each $j \in \{1, \ldots, p\}$, $|\big(X^tr\big)_j| \leq \lambda_p.\sigma$, 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 parameters to zero i.e they perform variable selection.

Remark 0.3.5. *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 provide with detailed supporting proof in [*[5], theorem 1.1; theorem 1.2*]

0.3.4 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 0.3.6. *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 high dimension setting where the model include tons more variables, far too many to know everything about ?.*

When you have million 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 don't not 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 $\lambda_1 \parallel \beta \parallel_1$ with the Ridge regression penalty $\lambda_2 \parallel \beta \parallel_2^2$. The Lagrangian problem become

$$minimize_{\beta}\{\parallel y - X\beta \parallel^2 + \lambda_1 \parallel \beta \parallel_1 + \lambda_2 \parallel \beta \parallel^2\}.$$

Altogether, Elastic-Net regression combines the strengths of Lasso and Ridge regression. Note that the Lasso and Ridge regression penalty get their own $\lambda's$; λ_1 for Lasso and λ_2 for Ridge.But more often, the problem is writing as

$$minimize_{\beta}\{\parallel y - X\beta \parallel^2 + \lambda(\alpha \parallel \beta \parallel_1 + (1-\alpha) \parallel \beta \parallel^2)\}, \quad for\ \alpha \in [0,1]\ and\ \lambda \geq 0$$

, say

$$\hat{\beta}^E(\lambda, \alpha) = argmin_{\beta}\{\parallel y - X\beta \parallel^2 + \lambda(\alpha \parallel \beta \parallel_1 + (1-\alpha) \parallel \beta \parallel^2)\}. \tag{29}$$

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 will control the mix between L_1 and L_2 regularization.

Remark 0.3.7. *We notice that:*

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

Cross-validation to find the best value of λ

There are various methods to select the "best" value for λ. One is to split the data into **K** chunks. We then use **K-1** of this 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 generalisable our model is.

We can use this new idea to choose a lambda value, by finding the lambda that minimises 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 $k = 1, \ldots, K$.The optimal λ is obtained by minimizing the total *Cross-validation* error:

$$\hat{\lambda} = \underset{\lambda}{argmin} \Big\{ \underbrace{\frac{1}{K} \sum_{k=1}^{K} \frac{1}{n_k} \parallel y_k - X_k \hat{\beta}_k(\lambda) \parallel_2^2}_{CV_{(K)}} \Big\}, \tag{30}$$

0.4 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 behaviour and to investigate if there was an difference in predictive power between the previous three regularization methods; ridge, Lasso and Elastic-net regression when they were applied on high-dimensional data.The statistical analysis was implement using **R** statistical software.

0.4.1 Simulated Data

For the simulation study, we use generalized linear model (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 [1] .The measures that are used to assess how good a logistic regression model is for prediction are : *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 tanking values between 0 and 1 (visit [2] for more details).The simulation study was inspired by the paper by *Krona* [2]. However, adjustments were made to the simulated datasets.

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.

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

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

were 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 equal the correlation matrix of X. Each predictor variable was assigned a predetermined $\beta - value$ .The response variable were simulated by running the simulated data through the inverse logit function (see [1]) ,

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

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.

Details information about the four examples :

Example 1 : we set the pairwise correlation between X_i *and* X_j predictors 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 in 8 groups, where the coefficients were set to pairwise be 0 and 2, $\beta = (\underbrace{2,2,\ldots,2}_{125},\underbrace{0,0,\ldots,0}_{125},\underbrace{2,2,\ldots,2}_{125},\ldots)^t$.

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, $\beta = (\underbrace{3, 3, \ldots, 3}_{500}, \underbrace{0, 0, \ldots, 0}_{500})^t$.

Results : The simulation of Example 1-4 was repeated 100 times: for every simulation, we calculated AUC, ME and their standard deviations (sd).In addition, the average number of selected variables by Lasso and the Elastic net was calculated.The results are summarized in table 1.

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

In example 2, the predictors were assigned coefficients β=0.8 with relatively high correlation amount predictors.As demonstrate in table 1, ridge regression improve over other methods considering AUC and ME. As mention in *subsection 1.3.1*, ridge regression tend to perform well under the circumstances in example 2.Moreover, the average number of coefficients for Lasso and elastic net was 20 and 328 respectively.In this setting, the elastic-net identify a larger number of coefficients that were correlated and non-zero.The Lasso, on the other hand results in a sparse final model but identify less of the 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 see that ridge regression outperform the Lasso and elastic net in view of the AUC.Since the elastic net and ridge regression perform considerably similar, they seem to perform equally as good in this setting.As discussed in earlier , ridge regression included all predictors in the final model and resulted in a less interpretable model.However, the elastic net identified on average 415 non-zero coefficients.supposedly, the elastic net adopted the grouping effect and correctly identified almost all non-zero coefficients simultaneously as it achieved high prediction accuracy.

In example 4, the predictors were divided into two groups of equal size that were assigned with β=3 and β=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 succeeded to identified approximately all non-zero coefficients as a result of the grouping effect.

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 tend to spread the coefficients shrinkage over a large number of coefficients, it achieve high predictive power throughout example 1-4.especially, the results in example 3 demonstrated the capacity of ridge regression.We identify that when the number of predictor are very large and a larger fraction of them most be included in the model, ridge regression dominates the Lasso and the elastic net.Consequently, it confirm that ridge regression is 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 to the quadratic penalty, while it simultaneously shrinks 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 elastic net approximately identified all-non zero coefficients in the simulations.In example 4, elastic-net performed grouped selection and showed to be a better variable selection method than

Lasso.Even though ridge regression did not incorporate variable selection, it achieved high prediction accuracy through-out example 1-4.Therefore, we observe that if the interpretability is not fundamental, ridge regression manage 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.

0.4.2 Real data example

Data description: For 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 normalised 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 year).This methylation object is a *"GenomicRatioset"*, a *Bioconductor*[19] data object derived from the *"SummarizedExperiment"*[15].These *"SummarizedExperiment"* objects contain *"assays"*, in this case normalised 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 *vignettes on Bioconductor*....url....

After reading in the data we can see in the provided R output that this object has $dim()\ of\ 5000 \times 37$, meaning it has 5000 features and 37 samples (observations).to extract the matrix of methylation M-values, we use "*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**.

Experiment steps : Let's denote by X the methylation matrix,

1) Singularity: we investigete singularity of the matrix $X^t X$ and check out what happen if we try to fit linear model to the data.

2) Ordinary least square versus Ridge regression : here, we work with a set of feature 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 extra the first 20 features of the features identified by Horvath, investigate correlations and we split the methylation data matrix and the age vector into training an 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 egularization 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.

Results.

1) We can see that we are able to get some effect size estimates, but they seem vert high.The *"Summary"* also says that we were unable to estimate effects sizes for 4964 features because of singularities.What this mean is that R couldn't find a way to perform the calculations necessary due to the fact that we have more features than observations.

2) Preditors are correlated 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 : $MSE_{lm} = 45.14 \ \geq \ MSE_{ridge} = 25.30$. *see also comments of Figures 1.4 and 1.5*

3) Comparing the feature selected by Lasso (41 features) and the elastic net (60 features) with Horvath signature, we can see that we selected some of the same feature (8 features for Lasso and 11 features for elastic net).*see also the comments of the remaining 6 figures.*

0.5 Measurement Error In Regression theory

In some sense, all statistical problems involve measurement error.

Measurement error occur whenever we cannot exactly observe one or more of the variables that enter into a model of interest.There are many reason such errors occur, the most common ones being "sampling error and instrument error".Where any notation is used here, the true value is denoted "X" and the variable observed in place of "X" 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 error occur in nearly every discipline; Here is a collection of examples in biomedical field:

Genomic: 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 statut: I n epidemiology , the outcome variable is often presence or absence of a disease such as breast cancer, hepatitis, AIDS. . . .This is often assessed through an imperfect diagnostic procedure such as an imaging technique or a blood text which can lead to either false positives or false negatives (misclassification).

Objective and some terminology

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

These are three general objective in measurement error problem we will try to address in this parts of our work.

The Model Description

One of the fundamental assumption in the linear regression analysis is that all observations are correctly observed.When this assumption is violated the measurement error creep into the data.The the usual statistical tools tend to loose their validity(*see [8] and [14] for more details.*). And important issue in the area of measurement errors is to find the consistent estimators of the parameters which can be accomplished by using some additional information from outside the sample.

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

$$y = X\beta + \epsilon \text{ , } W = X + U \tag{31}$$

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

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

For the sake of notation simplicity, we assume that $\beta_0 = 0$. The true covariate X are not observed, and instead we have noisy measurements $W = X + U$ where U is and $n \times p$ 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 $k - th$ variables, $\Sigma_{U(k,k)} = 0$.

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 0.5.1. *It follow from the structural model*

$$y_i = \beta^t X_i + \epsilon_i,\ W_i = X_i + U_i \tag{32}$$

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,*

$$\Gamma = \begin{bmatrix} \sigma_Y^2 & \Sigma_{YW} \\ \Sigma_{WY} & \Sigma_W \end{bmatrix} = \begin{bmatrix} \sigma^2 + \beta^t \Sigma_X \beta & \beta^t \Sigma_X \\ \Sigma_X \beta & \Sigma_X + \Sigma_U \end{bmatrix}. \tag{33}$$

This lead to:

$$y_i | W_i = w_i = \gamma^t w_i + \delta_i \tag{34}$$

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

Theorem 0.5.1. *Under the given assumptions,* γ *and* σ_δ^2 *are given by,*

$$\gamma = \left(\Sigma_W\right)^{-1} \Sigma_X \beta = \left(\Sigma_X + \Sigma_U\right)^{-1} \Sigma_X \beta \tag{35}$$

$$\sigma_\delta^2 = \sigma^2 + \beta^t \Sigma_X \beta - \gamma^t \left(\Sigma_X + \Sigma_U\right) \gamma \tag{36}$$

Thus

$$\beta = \mathcal{K}_X^{-1} \gamma. \tag{37}$$

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) [16] and Aickin and Ritenbaugh (1992)* for example,discussion and illustrations of the role of reliability matrix.

Estimated Coefficients and Behaviour 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 $\hat{\beta} = (X^t X)^{-1} X^t y$ (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 0.5.1. *The maximum likelihood estimators of* γ *and* σ_δ^2 *are just the naive least squares estimators,*

$$\hat{\beta}_{naive} = \hat{\gamma} = (W^t W)^{-1} W^t y = S_{WW}^{-1} S_{Wy}\ ,\ \hat{\sigma}_{naive}^2 = \hat{\sigma}_{delta} = \frac{1}{n-p} \sum_i (y - \hat{y}_i)^2,\ with\ \hat{y}_i = \hat{\beta}_{naive}^t w_i \tag{38}$$

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

Proposition 0.5.2. *The exact bias expression for the naive estimators under the given assumptions are given by:*

$$\mathbb{E}\left[\hat{\beta}_{naive}\right] = \gamma = \mathcal{K}_X\beta \;,\; \mathbb{E}\left[\hat{\sigma}^2_{naive}\right] = \sigma^2_\delta \tag{39}$$

Remark 0.5.2. *This result lead 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.*

Correcting for Measurement Error in Multilinear regression

With some exceptions (*see [4], chap11 and 12*), correcting for measurement error requires informations or data as laid out in item **3)** section 0.5.

Myriad approaches to carrying out corrections for measurement error have emerged,A number of which are described in [4].These include *direct bias correction, moment based approach, likelihood based techniques,SIMEX and techniques based on modifying equations.*

Proposition 0.5.3. *When Σ_U is known and $\mathcal{K}_X$ is unknown, then $\mathcal{K}_X$ is estimated consistently by replacing Σ_X and $\Sigma_W$by their respective consistent estimators as:*

$$\hat{\Sigma}_X = \hat{\Sigma}_W - \Sigma_U \;,\; \hat{\Sigma}_W = S_{WW} = \frac{W^tW}{n}; \quad \textit{and we have } \hat{\mathcal{K}}_X = S^{-1}_{WW}\left(S_{WW} - \Sigma_U\right). \tag{40}$$

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

$$\hat{\beta} = \hat{\mathcal{K}}^{-1}_X\hat{\gamma} = \left(S_{WW} - \Sigma_U\right)^{-1} S_{Wy}, \;,\; \hat{\sigma}^2 = \hat{\sigma}^2_\delta - \hat{\beta}^t\Sigma_U\hat{\mathcal{K}}_X\hat{\beta} \tag{41}$$

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

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

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

Proposition 0.5.4. *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:*

$$\hat{\Sigma}_U = \frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{k=1}^{m_i}\left(W_{ik} - \bar{W}_{i.}\right)\left(W_{ik} - \bar{W}_{i.}\right)^t}{m_i - 1} \tag{42}$$

In that case;

$$\hat{\Sigma}_X = S_{WW} - \hat{\Sigma}_U,\ \hat{\mathcal{K}}_X = S_{WW}^{-1}\left(S_{WW} - \hat{\Sigma}_U\right),\ and\ \hat{\beta} = \left(S_{WW} - \hat{\Sigma}_U\right)^{-1} S_{Wy} \tag{43}$$

Remark 0.5.3. *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 [14]* who wrote the first influential book focusing on linear regression models, and on by *Caroll et al. [7]* who treated measurement error in a much broader application context.Another book that give wide treatment to the topic is by *Buanaccorsi [8]* who focuses on different topics from those in the aforementioned two books and places emphasis on more applied approach.

0.6 Measurement Error in High-Dimensional Context :Behaviour and Correction Methods

0.6.1 Ridge Regression Estimation Over Measurement Error Ridden Data.

The standard assumption in the linear regression analysis is that explanatory variables are uncorrelated.When this assumption is violated, the explanatory variables are nearly dependent, which refers as **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 [18]* was discuss in section 0.3.1 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 , then and important issues is how to obtain the consistent estimators of regression coefficients.One simple idea is to use the ridge regression estimation over the error ridden data.An obvious question that crops up is what happen then?.

In this section, we attempt to answer such questions.

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 31 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 :

$$\hat{\beta}^{LS}_{ME} = \mathcal{K}_X^{-1}\hat{\gamma} = \left(S_{WW} - \Sigma_U\right)^{-1} S_{Wy} \text{ (41)}, \; and \; \gamma = \mathcal{K}_X\beta \text{ (35)}$$

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

$$\underset{\beta\in\mathbb{R}^p}{minimize}\left\{ \parallel y - W\gamma \parallel_2^2 \right\} \quad subject\ to\ \parallel \beta \parallel^2 \leq s\ for\ some\ constant\ s \tag{44}$$

the Lagrangian problem become

$$\underset{\beta\in\mathbb{R}^p}{minimize}\left\{ \parallel y - W\mathcal{K}_X\beta \parallel_2^2 + k \parallel \beta \parallel^2 \right\} \tag{45}$$

Proposition 0.6.1. *The numerical solution of this problem corresponding to the ridge regression estimator of β in measurement error model 31 is given by:*

$$\hat{\beta}^{R}_{ME} = \left[\mathbb{I}_p + k\left(n\mathcal{K}_X^t S_{WW}\mathcal{K}_X\right)^{-1}\right]^{-1} \hat{\beta}^{LS}_{ME}. \tag{46}$$

Corollary 4. *Substituting the consistent estimator of $\mathcal{K}_X$ given in* (40) *we get,*

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

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:*

$$MSE\left(\hat{\beta}^{R}_{ME}, k\right) = k^2\beta^t\left[C_n + k\mathbb{I}_p\right]^{-2}\beta + \sigma_\delta^2 tr\left(Z_n^{ME} C_n^{-1} \left(Z_n^{ME}\right)^t\right) \tag{48}$$

Remark 0.6.1.

- *When $n \to \infty$ then $C_n \to C$, $Z_n^{ME} \to Z^{ME}$ and*

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

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

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, $P^tCP = D = diag(\lambda_1, \ldots, \lambda_p)$ (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.

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

and

$$\sigma_\delta^2 tr\left(Z^{ME}C^{-1}\left(Z^{ME}\right)^t\right) = \sigma_\delta^2\sum_{j=1}^{p}\frac{\lambda_j}{(\lambda_j+k)^2}\ , \ \textit{see } \textbf{Remark 1.2.1 and (1.19)}$$

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

$$MSE\left(\hat{\beta}^R_{ME}, k\right) = k^2\sum_{j=1}^{p}\frac{\alpha_j^2}{(\lambda_j+k)^2} + \sigma_\delta^2\sum_{j=1}^{p}\frac{\lambda_j}{(\lambda_j+k)^2} = \psi_b(k) + \psi_v(k). \tag{49}$$

Theorem 0.6.1 (*from [21]*)**.** *There always exist a $k > 0$ such that ,*

$$MSE\left(\hat{\beta}^R_{ME}, k\right) < MSE\left(\hat{\beta}^{LS}_{ME}\right)\ . \tag{50}$$

0.7 Measurement Error In Lasso

Modern statistics is facing problems due to the increase of dimensionality of the data in field such as genomics,finance,network analysis,. . . .It is quite canonical 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 has 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 is available, among which the "Lasso regression [23], "Dantzig selector (DS) [8] and Smoothly Clipped Absolute Deviation (SCAD) [14].These methods all allow model selection and parameter estimation through a penalization of the parameters as seen for the Lasso case.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 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 (39) and to loss of power [7].

Since the standard Lasso is widely used despite the present of measurement error, it is of interest to study the effects measurement error has on the analysis and describes some of the statistical methods used to correct for those effects.

0.7.1 Impact Of Ignoring Measurement Error

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

$$S_{WW} = \begin{bmatrix} S_{WW}(S,S) & S_{WW}(S,S^c) \\ S_{WW}(S^c,S) & S_{WW}(S^c,S^c) \end{bmatrix} \tag{51}$$

We saw that in the absence of measurement error, the Lasso is consistent for prediction and estimation (22)). $y = X\beta + \epsilon = (W+U)\beta + \epsilon = W\beta + \underbrace{\epsilon - U\beta}_{\delta}$.

Proposition 0.7.1. *Assume the compatibility condition* (21) *holds with constant* Φ *, and that there exist a constant* λ_0 *such that* $\frac{2}{n} \parallel \delta^t W \parallel_\infty \leq \lambda_0$*; Then, with a regularization parameter* $\lambda \geq 2\lambda_0$*,*

$$\frac{1}{n} \parallel W(\hat{\beta}^{Lasso} - \beta) \parallel_2^2 + \lambda \parallel \hat{\beta}^{Lasso} - \beta \parallel_1 \leq \frac{4\lambda^2 s}{\Phi_o^2}. \tag{52}$$

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

$$\parallel \delta^t W \parallel_\infty \leq \parallel \epsilon^t W \parallel_\infty + \parallel \beta^t U^t X \parallel_\infty + \parallel \beta \parallel_1 \parallel U^t U \parallel_\infty$$

Hence the bound (52) is implied by ,

$$\frac{2}{n} \parallel \epsilon^t W \parallel_\infty + \frac{2}{n} \parallel \beta^t U^t X \parallel_\infty + 2 \parallel \beta \parallel_1 \parallel \frac{U^t U}{n} \parallel_\infty \leq \lambda_0\ ; \tag{53}$$

and the Lasso with measurement error is consistent if all the three terms in the above expression (53) converge to 0.However,

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

, consequently, we do not obtain consistency.

We have just see 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 affect covariate selection with Lasso.By definition (27), the "irrepresentable condition with measurement error (**IC-ME**) hold if there exists a constant $\theta \in [0,1[$ such that ,

$$\parallel S_{WW}(S^c,S) S_{WW}(S,S)^{-1} sign(\beta_S) \parallel_\infty \leq \theta.. \tag{54}$$

In presence of measurement error, *Sorensen, Frigessi and Thoren (2015) [22]* shown that to achieve covariate selection consistency, we need the following additional condition called "Measurement Error Condition" (**MEC**):

Definition 0.7.1 (MEC)**.** *The measurement error condition (MEC) is satisfied if*

$$\Sigma_W(S^c,S)\Sigma_W(S,S)^{-1}\Sigma_U(S,S)-\Sigma_U(S^c,S)=0.\,,\ (visit\ [22]\ for\ more\ details). \tag{55}$$

0.7.2 Correction for Measurement Error in Lasso

The purpose of this section es to describe some penalized regressions correction methods that may be used to correct both the variable selection and the model estimation at the same time assuming measurement error is adequately modelled (in our case " additive measurement error").

To show the bias in the estimation caused by measurement error, consider the naive Lasso approach, plugging in W for X in the Lasso estimator defined in (16)

$$\hat{\beta}^{LS}(\lambda_n)=argmin_{\beta\in\mathbb{R}^p}\Big\{\parallel y-W\beta\parallel_2^2+\lambda_n\parallel\beta\parallel_1\Big\}. \tag{56}$$

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

$$\mathbb{E}\big[\parallel y-W\beta\parallel_2^2|X,y\big]=\parallel y-X\beta\parallel_2^2+n\beta^t\Sigma_U\beta. \tag{57}$$

Corrected Lasso (Non Convex Lasso)

The must natural way for correcting for the bias in (57) leads to the constrained correct Lasso (CCL):

$$\hat{\beta}_{CCL}\in\underset{\beta:\|\beta\|_1\leq R}{argmin}\Big\{\frac{1}{n}\parallel y-W\beta\parallel_2^2-\beta^t\Sigma_U\beta\Big\}. \tag{58}$$

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

$$\hat{\beta}_{RCL}\in\underset{\beta\in\mathbb{R}^p}{argmin}\Big\{\frac{1}{n}\parallel y-W\beta\parallel_2^2-\beta^t\Sigma_U\beta+\lambda_{RCL}\parallel\beta\parallel_1\Big\}. \tag{59}$$

both introduced by *Loh and Wainright (2012) [22].*

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

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

. Notice that Σ_U is in practice unknown and must be estimated from data.

Proposition 0.7.2. *The estimator* (58) *and* (59) *can be reformulated as:*

$$\hat{\beta}_{CCL} \in \underset{\beta:\|\beta\|_1 \leq R}{argmin} \left\{ \frac{1}{2}\beta^t \hat{\Sigma}_X \beta - \hat{\gamma}^t \beta \right\}., \text{ and} \tag{60}$$

$$\hat{\beta}_{RCL} \in \underset{\beta:\|\beta\|_1 \leq b_0\sqrt{s}}{argmin} \left\{ \frac{1}{2}\beta^t \hat{\Sigma}_X \beta - \hat{\gamma}^t \beta + \lambda_{RCL} \parallel \beta \parallel_1 \right\}., \text{ for some constant } b_0. \tag{61}$$

Remark 0.7.1. *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 semidefinite 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* (58) *and* (59) *are **non convex**.*

Remark 0.7.2. *When, $\hat{\Sigma}_X$ has negative eigenvalues (which happen very often under high-dimensionality), the objective function in equation* (59) *is unbounded from below, hence we make use of the regularized estimator* (61) *to overcome these technical difficulties.*

Remark 0.7.3. *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 [11] demonstrated that a simple "**project gradient descent algorithm**" applied to the problems* (58) *or* (61) *(if b_0 is properly chosen) converge with high probability to a small neighbourhood of the set of all global minimizers.*

Definition 0.7.2. *Project gradient descent is a standard way to solve constrained optimization problem.*

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 $G > 0$ (≥ 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 the 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,

$$\left(G\right)_+ = \underset{G_1 \geq 0}{argmin} \parallel G - G_1 \parallel_{\max} \tag{62}$$

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

$$\hat{\beta}_{CoCo} = \underset{\beta \in \mathbb{R}^p}{argmin} \left\{ \frac{1}{2} \beta^t \tilde{\Sigma}_X \beta - \hat{\gamma}^t \beta + \lambda_{CoCo} \parallel \beta \parallel_1 \right\} \tag{63}$$

Remark 0.7.4. *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* (63) *is guaranteed to be convex.*

Theorem 0.7.1 (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,*

$$A = L^T L \tag{64}$$

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 (63) can be reformulates as:

$$\hat{\beta}_{CoCo} = \underset{\beta \in \mathbb{R}^p}{argmin} \left\{ \frac{1}{n} \parallel \tilde{y} - \tilde{X}\beta \parallel_2^2 + \lambda_{CoCo} \parallel \beta \parallel_1 \right\} \tag{65}$$

Remark 0.7.5. *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 [12] or Least angle regression [10] to obtain solution.*

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 ,

$$CV_{(K)} = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{n_k} \parallel y_k - W_k \hat{\beta}_k(\lambda) \parallel_2^2 . \tag{66}$$

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 (30) in the same way we shown that the loss function in (56) is a biased version of the one in (16).Observing that (30) is equivalent to:

$$\hat{\lambda} = \underset{\lambda}{argmin}\left\{ \frac{1}{K}\sum_{k=1}^{K}\frac{1}{2}\hat{\beta}_k^t(\lambda)\Sigma_k\hat{\beta}_k(\lambda) - \gamma_k^t\hat{\beta}_k(\lambda) \right\}. \tag{67}$$

where $\Sigma_k = \frac{1}{n_k}X_k^tX_k$ and $\gamma_k = \frac{1}{n_k}X_k^ty_k$.

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 [9]* substituted Σ_k and γ_k with their projected and estimated counterparts $\tilde{\Sigma}_k = \left(\hat{\Sigma}_k\right)_+$ and $\hat{\gamma}_k$ .With this correction, the cross-validated λ is defined as:

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

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

0.8 Matrix uncertainty selector (MU-Selector)

So far, we saw that corrected Lasso (NCL) (61) and CoCoLassso correct for measurement error, by including in the model the covariance of the measurement error Σ_U, and yielding estimators with good theoretical properties.However, this quantity is assumed to be known and in practice it is usually not known.The estimation of the covariance matrix of the measurement error requires additional data as replicated measurement 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 [20].*

We consider the model in (31).We typically assume that β is "s-sparse" where $1 \leq s \leq p$ is some integer. In what follows, we assume that ϵ and U satisfy the assumptions:

$$\frac{1}{n} \parallel W^t\epsilon \parallel_\infty \leq \lambda \ \ and \ \ \parallel U \parallel_\infty \leq \delta. \quad (\ with\ high\ probability\ ). \tag{69}$$

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

$$\min\left\{ \parallel \beta \parallel_1 : \beta \in \Theta, \frac{1}{n} \parallel W^t(y - W\beta) \parallel_\infty \leq (1+\delta)\delta \parallel \beta \parallel_1 + \lambda \right\}, \tag{70}$$

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

The problem (70) 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 $\frac{1}{n}X^tX$ are equal to 1.

Proposition 0.8.1 (solution existence). *Under assumptions* (69) *, the feasible set of the convex problem* (70) *is non empty,*

$$\Psi = \left\{ \beta \in \Theta, \frac{1}{n} \parallel W^t(y - W\beta) \parallel_\infty \leq (1+\delta)\delta \parallel \beta \parallel_1 + \lambda \right\} \neq \emptyset \tag{71}$$

Remark 0.8.1. *If* $\delta = 0$ *and* $\Theta = \mathbb{R}^p$ *, the MU-Selector becomes the Dantzig selector* (28). *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").*

0.9 Numerical Study

0.9.1 Ridge under measurement error (simulation)

As discuss earlier, ridge regression (13) provide better estimators when facing problem of multicollinearity in our data.The purpose of this simulation is to evaluate the performance of the modified ridge estimation in (45) when problem of multicollinearity is present in the measurement error ridden data.To this end, we will restrict particularly to the case where p<n (low-dimensional data) with high correlations between covariates measured with error.

Simulation design: We simulate data from the true model ,

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

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 ,

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

. The simulated data was divided into a training and a test set.The four methods ;*True OLS*[5] *($y \sim X$)*(4) , *corrected OLS*(41) , *naive ridge* and *modified ridge regression* (45) were used to fit a corresponding model to the training set. The fitted models were used to make predictions to the test set and we computed the MSE and the PE (prediction error on the test set).The procedure was repeated 100 times.

[5] Ordinary least square

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 three others.Meaning that the provided $\hat{\beta}$ is much more reliable considering MSE (as mentioned in *theorem 3.1.2*) and PE .We also find out in passing that using the corrected version of OLS (41) in this setting (*"high-correlation with measurement error"*) would result to a pretty poor estimator given the MSE and PE (table 2).

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
	Example 3			Example 4		
models	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.)*

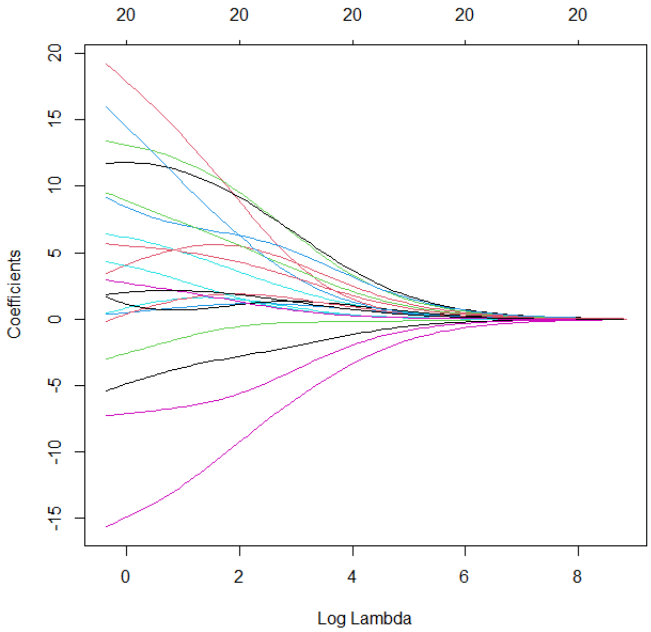


Figure 2: *plot showing how estimated coefficients for each methylated site change as we increase the penalty* λ. *We can see that initially, some parameter estimates are really large, and these tend to shrink fairly rapidly.*

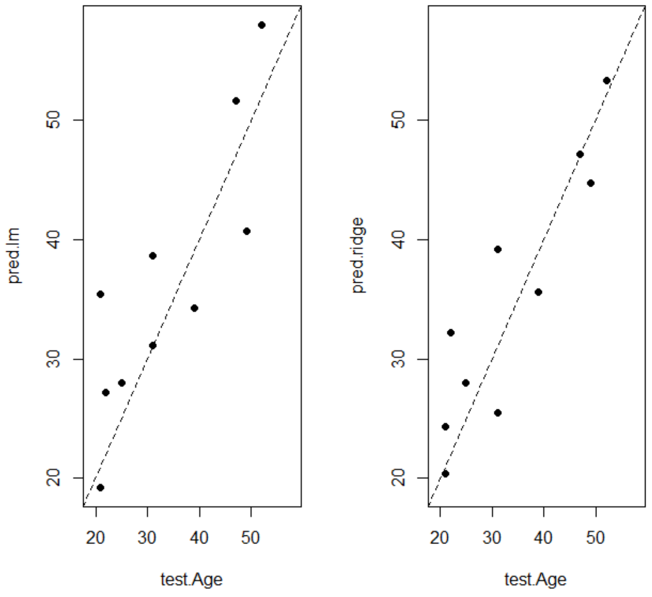


Figure 3: *Predicted Ages for each methods against the true Ages. The ridge ones are much less spread out with far fewer extreme predictions.*

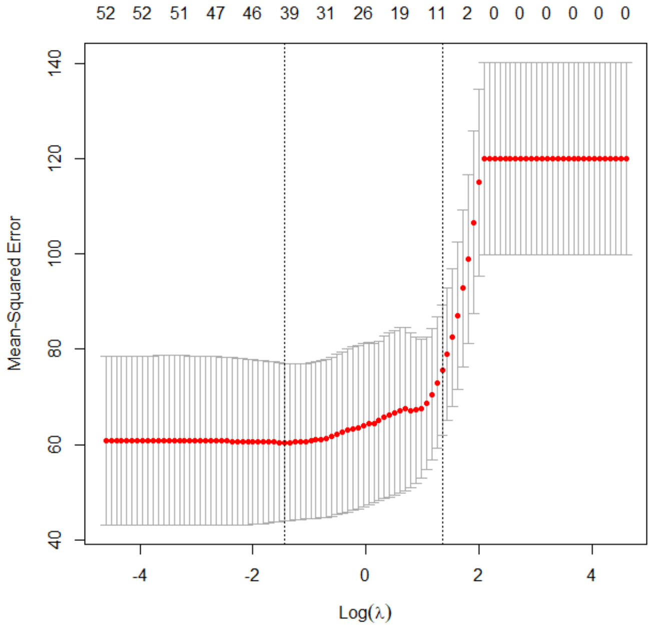


Figure 4: *Cross-validation performance for Lasso.*

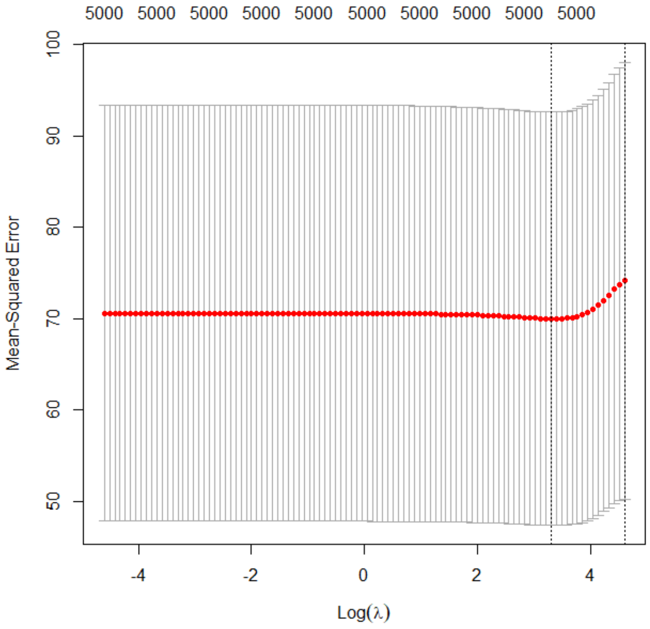


Figure 5: *Cross-validation performance for Ridge.*

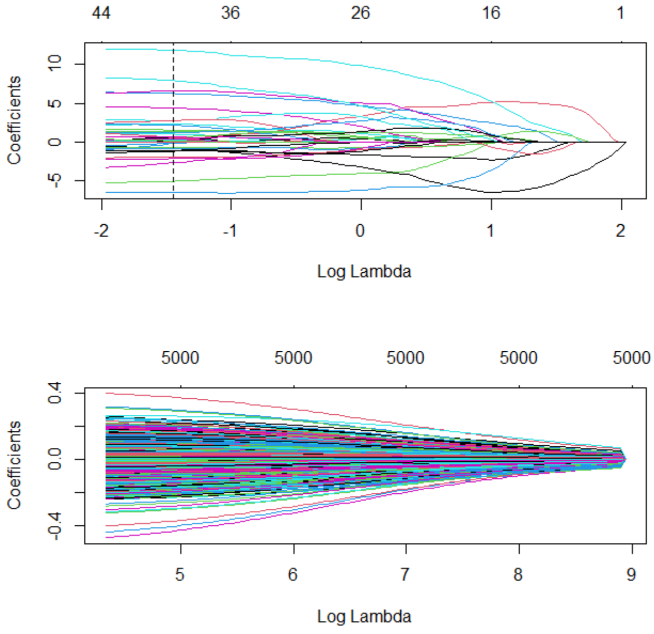


Figure 6: *The paths tend to go exactly to zero much more when sparsity increases when we use lasso model.In ridge case, the paths tends toward zero but less commonly reach exactly zero.*

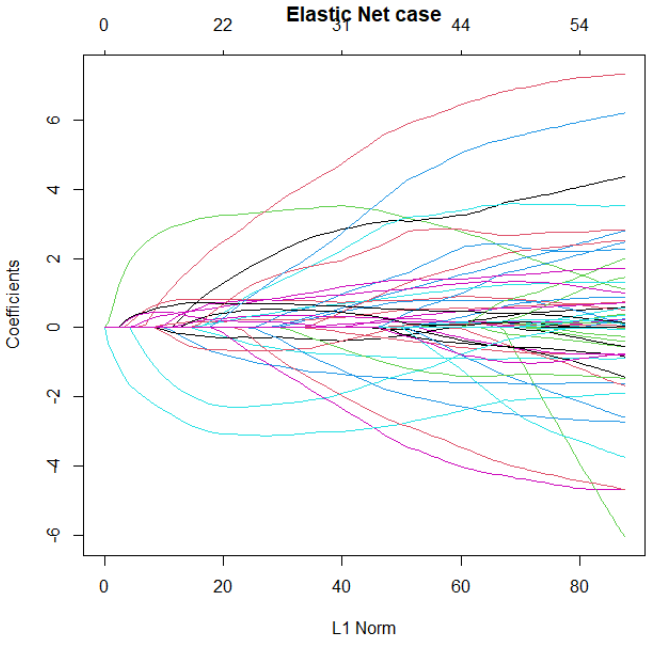


Figure 7: *Coefficients paths elastic net.We can see that coefficients tend to go exactly to zero, but the paths are a bit less extreme than with pure Lasso; similar to ridge.*

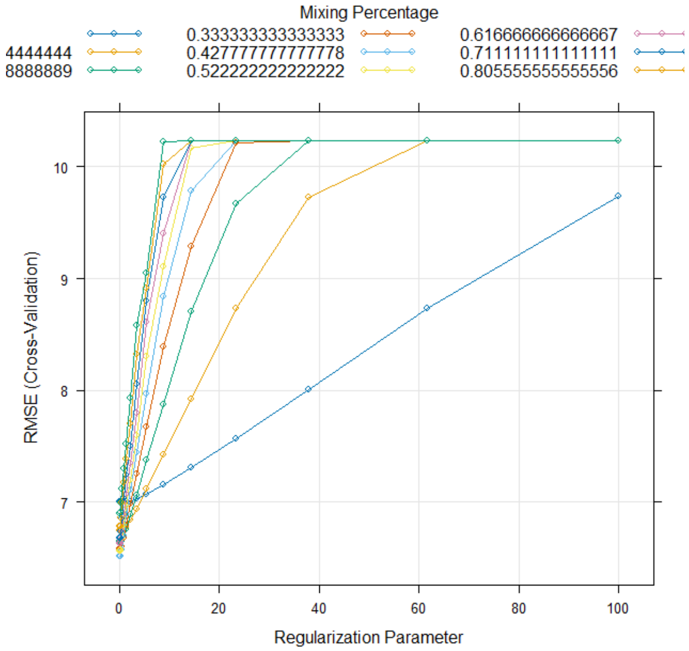


Figure 8: *Cross-validation to find the optimal pair of* (α, λ) *for elastic net (mixing percentage).*

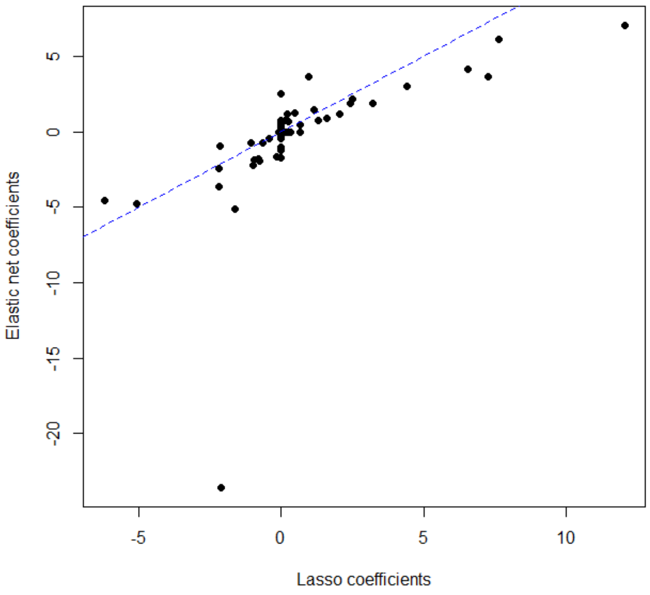


Figure 9: *Lasso coefficients against elastic net coefficients.We can see that the coefficients from these two methods are broadly similar, but the elastic net coefficients are a bit more conservative.*

	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.)*

Bibliography

[1] A. Agresti. Categorical data analysis, volume 792. John Wiley & Sons, 2012.

[2] A. J. Bowers and X. Zhou. Receiver operating characteristic (roc) area under the curve (auc): A diagnostic measure for evaluating the accuracy of predictors of education outcomes. Journal of Education for Students Placed at Risk (JESPAR), 24(1):20–46, 2019.

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

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

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

[6] 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.

[7] 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.

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

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

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

[11] 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.

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

[13] J. Friedman, T. Hastie, R. Tibshirani, B. Narasimhan, K. Tay, N. Simon, and J. Qian. Package âglmnetâ. CRAN R Repositary, 2021.

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

[15] L. Geistlinger, G. Csaba, M. Santarelli, M. Ramos, L. Waldron, R. Zimmer, M. L. Geistlinger, D. SummarizedExperiment, G. GSEABase, K. KEGGREST, et al. Package âenrichmentbrowserâ. 2016.

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

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

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

[19] M. Ramos, L. Schiffer, A. Re, R. Azhar, A. Basunia, C. Rodriguez, T. Chan, P. Chapman, S. R. Davis, D. Gomez-Cabrero, et al. Software for the integration of multiomics experiments in bioconductor. Cancer research, 77(21):e39–e42, 2017.

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

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

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

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

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