Al-Qadisiyah Journal for Administrative and Economic Sciences ISSN$_{Online}$: 2312-9883
QJAE, Volume 26, Issue 3 (2024) ISSN$_{Print}$: 1816-9171

Bayesian Ridge and Reciprocal Ridge: A Comparative Study

Rahim Alhamzawi
Muna Qassim Shemran
University of Al-Qadissyia

Corresponding Author : *Muna Qassim Shemran*

Abstract:In the last fifty years, considerable efforts have been made into the development of ridge regression model, which employs a symmetric penalty function about 0,continuous and non-decreasing in (0, ∞). The reciprocal ridge regression model is an extension of the ridge regression, which employs a decreasing penalty function that is continuous at 0 and converges to infinity when the regression coefficients become closer to zero in the interval (0, ∞). It is well known that the ridge regression is a special case from bridge regression (Frank and Friedman, 1993), whereas the reciprocal ridge regression is a special case from reciprocal bridge regression (Song and Liang, 2015).
The objective of this thesis is to examine if the Bayesian reciprocal ridge regression can provide better performance in prediction than Bayesian ridge regression when handling multicollinearity and high-dimensional problems. Simulation results and a popular air pollution data analyses show that both approaches have good mixingproperties and perform comparably to each to other in terms of prediction accuracy.

Introduction: The linear regression model is extensively utilized in various fields such as agriculture, economics, ecology, engineering, finances, and medical investigations.
It is utilized to model the correlation between a desired outcome and a group of factors.
Let's examine the linear regression model:

$$y = X\beta + \epsilon \qquad 1.1$$

where y = (y1, · · · , yn) is the n × 1 vector of centered outcomes, X=(x1, · · · , xn)j is the n × p matrix of standardized covariates, β = (β1, · · · , βp)j is the p × 1 regression coefficients vector to be estimated and s = (ϵ1, · · · , ϵn) is the n × 1 vector of i.i.d Gaussian (normal) errors with mean 0 and variance σ2, i.e. ϵi ∼ N (0, σ2) for i = 1, · · · , n.
The regression coefficients vector β can be estimated using the Ordinary Least Squares (OLS) estimator, which is obtained by solving the following problem:

$$\min_{\beta} (y - X\beta)'(y - X\beta). \qquad 1.2$$

The OLS estimator (classical estimator) is given by:

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'y. \qquad 1.3$$

The above estimator coincides with the maximum likelihood estimator (MLE) Hastie et al. (2009). From (1.3), it can be observed that βˆOLS solution depends on the non-singularity of (XjX). If (XjX) is nonsingular, then we can obtain unique solution of βˆOLS . If (XjX) is not singular, then the unique βˆOLS solution cannot be estimated. In the case of high dimensional data (p n) or in the case of presence of multicollinearity, it may be difficult to find the approximate inverse of (XjX). The βˆOLS can be shown to be unbiased as follows (Lancaster, 1990):

$$\begin{aligned}\mathbb{E}(\hat{\beta}_{OLS}) &= \mathbb{E}[(X'X)^{-1}X'y] \\ &= (X'X)^{-1}X'\mathbb{E}[y] \\ &= (X'X)^{-1}X'\mathbb{E}[X\beta+\epsilon] \\ &= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'\mathbb{E}[\epsilon] \\ &= \beta\end{aligned} \quad 1.4$$

The variance-covariance matrix of $\hat{\beta}OLS$ can be estimated as follows.

$$\begin{aligned}Cov(\hat{\beta}_{OLS}) &= Cov[(X'X)^{-1}X'y] \\ &= (X'X)^{-1}X'Cov[y]X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}\end{aligned} \quad 1.5$$

From (1.4) and (1.5) we have:

$$\hat{\beta}_{OLS} \sim N(\beta, \sigma^2(X'X)^{-1}). \quad 1.6$$

Under ordinary least squares (OLS) assumptions, OLS estimator is BLUE (best linear unbiased estimator). Therefore, it is the best (efficient) estimator. Nonetheless, there are estimators of β with smaller variance that have some bias.

$$MSE(\hat{\beta}_{OLS}) = \mathbb{E}(\hat{\beta}_{OLS}-\beta)^2 = Var(\hat{\beta}_{OLS}) + \underbrace{(E(\hat{\beta}_{OLS})-\beta)^2}_{=(bais\ in\ \hat{\beta}_{OLS})^2}. \quad 1.7$$

Classical Ridge Regression

Ridge regression is an OLS estimation, with a constraint on the sum of the squared coefficients. The estimate arises from solving the linear regression equation with a regularization regression parameter (tuning parameter) $\lambda > 0$ according to Hoerl and Kennard (1970), that is

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{Ridge}} = \min_{\boldsymbol{\beta}} (\boldsymbol{y}-\boldsymbol{X\beta})'(\boldsymbol{y}-\boldsymbol{X\beta}) + \lambda\sum_{j=1}^{p}\beta_j^2. \quad 1.8$$

The second term in 2.1 is called a shrinkage penalty. When the regularization parameter $\lambda > 0$, the shrinkage penalty has no effect, and ridge regression will produce the OLS estimates. When the regularization parameter $\lambda \to \infty$, the impact of the shrinkage penalty grows, and the $\hat{\boldsymbol{\beta}}_{\boldsymbol{Ridge}}$ will approach zero. The Ridge estimator is given by:

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{Ridge}} = (X'X+\lambda I)^{-1}X'\boldsymbol{y}, \quad 1.9$$

where $\boldsymbol{I}$ is the identity matrix. The term $\lambda\boldsymbol{I}$ make the variance covariance matrix $(\boldsymbol{X}^{\prime}\boldsymbol{X}+\lambda\boldsymbol{I})$ is always invertible and produces a unique solution of $\hat{\boldsymbol{\beta}}_{\boldsymbol{Ridge}}$ even if $(\boldsymbol{X}^{\prime}\boldsymbol{X})$ singular. Thus, $\hat{\boldsymbol{\beta}}_{\boldsymbol{Ridge}}$ are much stable compared to the $\hat{\boldsymbol{\beta}}_{\boldsymbol{OLS}}$ estimates which may be unstable when covariates are correlated or in case p n. The above estimator (2.2) gives very better accuracy in predicting the model but it creates challenge in interpreting the model when the number of covariates is very large. This technique shrinks the regression coefficients towards 0, but never sets them actually equal to 0.The above estimator can be written as (Hoerl and Kennard, 1970; Hoerl et al., 1975):

$$\hat{\boldsymbol{\beta}}_{Ridge} = (X'X + \lambda I)^{-1}X'X(X'X)^{-1}X'\boldsymbol{y} = (I + \lambda(X'X)^{-1})^{-1}\hat{\boldsymbol{\beta}}_{OLS}. \quad 1.9$$

This shows that $\hat{\boldsymbol{\beta}}_{Ridge}$ is a linear combination of the $\hat{\boldsymbol{\beta}}_{OLS}$. The key properties with Ridge estimator are:

$$\mathbb{E}[\hat{\boldsymbol{\beta}}_{Ridge}] = (X'X + \lambda I)^{-1}(X'X)\boldsymbol{\beta}, \quad 1.10$$

and

$$Var[\hat{\boldsymbol{\beta}}_{Ridge}] = \sigma^2(X'X + \lambda I)^{-1}(X'X)(X'X + \lambda I)^{-1}. \quad 1.11$$

We use the R function genet (x, y, alpha=0) to estimate the ridge estimators. Setting alpha=0 is equivalent to ridge regression, setting alpha=1 is equivalent to lasso regression and if 0<alpha<1 is equivalent to elastic net regression.

Bayesian estimation for ridge regression

The instability of the ordinary least squares (OLS) estimator is widely recognised when multicollinearity is observed. Furthermore, when k is not equal to n, it is recognised that the estimator produced is non-unique due to X having less than full rank. In order to enhance the predictive precision of ordinary least squares (OLS), other regression; techniques incorporating different forms of penalties have been devised. Ridge regression, introduced by Hoerl and Kennard in 1970, is a statistical method that involves shrinking the regression coefficients. This shrinkage is achieved by utilizing independent and identical normal priors with a mean of zero. It is employed to ensure stability in situations with substantial variability, such as regression with collinearity $p>n$.

In ridge regression, the estimate of β is obtained using L2-constrained least squares،

$$\hat{\boldsymbol{\beta}} = \underbrace{\arg\min}_{\boldsymbol{\beta}}(\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta}) + \lambda||\boldsymbol{\beta}||^2, \quad (2.11)$$

where $||\boldsymbol{\beta}||^2 = \sum_{j=1}^{k} \beta_j^2$ and $\lambda \geq 0$ is a Lagrange multiplier which controls the degree of penalization. The ridge regression can be interpreted as a Bayesian posterior mode estimate when assigning an independent normal prior to each β_j, $P(\beta_j) = (2\pi\lambda^{-1}\sigma^2)^{-1/2}\exp\{-\frac{\lambda}{2\sigma^2}\beta_j^2\}$.

MCMC sampling for the Bayesian ridge regression

The joint posterior in (2.12) Provide us with a precise Gibbs sampler algorithm that commences with initial estimations for β, σ2, and λ and proceeds through the following steps:

1. Generate $\boldsymbol{\beta}$ from a multivariate normal distribution with mean $\boldsymbol{\Sigma}^{-1}\boldsymbol{X}'\boldsymbol{y}$ and variance covariance matrix $\sigma^2\boldsymbol{\Sigma}^{-1}$.

2. Generate σ^2 from an inverse Gamma distribution with shape parameter $n/2+k/2$ and scale parameter $\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})-\frac{\lambda}{2}\boldsymbol{\beta}'\boldsymbol{\beta}$.

3. Generate λ from a Gamma distribution with shape parameter $k/2+a$ and rate parameter $[\frac{\boldsymbol{\beta}'\boldsymbol{\beta}}{2\sigma^2}+b]$, where $a=b=0.1$.

Reciprocal Ridge Regression

In the last fifty years, considerable efforts have been made into the development of ridge regression model, which is employ symmetric penalty function about 0,continuous and nondecreasing in (0, ∞). The reciprocal ridge regression model is an extension of ridge regression, which employs a decreasing penalty function in (0, ∞), is discontinuous at 0 and tends towards infinity as the regression coefficients approach zero. It is well known that the ridge regression is a special case from bridge regression (Frank and Friedman, 1993), whereas the reciprocal ridge regression is a special case from reciprocal bridge regression (Song and Liang, 2015).

The reciprocal Ridge (rRidge) regression Utilizes a penalty function that decreases, as opposed to the Ridge regression strategy which applies increasing penalties on the regression parameters, resulting in a more robust regularization model. (Song and Liang) (2015) and Mallick et al. (2021)):

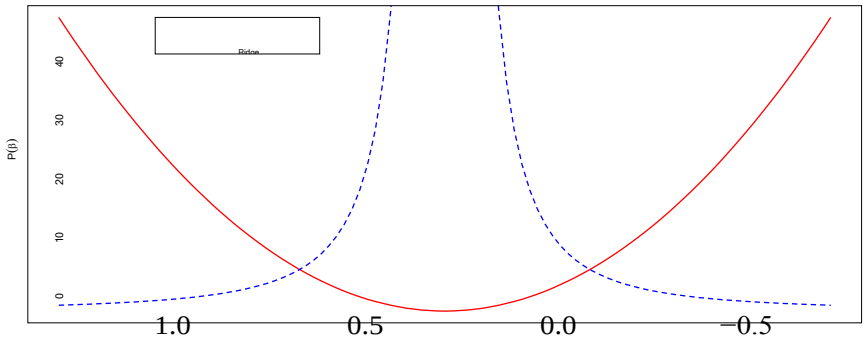

Figure 1: Comparison of shapes of ridge and rRidge penalty functions when λ = 1

$$\hat{\boldsymbol{\beta}}_{rRidge}=\min_{\boldsymbol{\beta}}\ (\boldsymbol{y}-X\boldsymbol{\beta})'(\boldsymbol{y}-X\boldsymbol{\beta})+\lambda\sum_{j=1}^{p}\frac{1}{\beta_j^2}I\{\beta_j\neq 0\}, \qquad 1.19$$

where $I(.)$ Represents a function that serves as an indicator and $\lambda > 0$ is the regularization parameter for the reciprocal ridge regression that controls the degree of penalization. The shapes of ridge and reciprocal ridge penalties are illustrated in Figures 3.1 and 3.2, which show that the reciprocal ridge give nearly zero coefficients infinity penalties; in contrast, ridge gives nearly zero coefficients nearly zero penalties. Solving the above optimization problem can be done by modifying the reciprocal lasso algorithm reported in (Song and Liang, 2015; Song, 2018) with the reciprocal ridge penalty using stochastic approximation annealing (Liang et al., 2014) algorithms.

Bayesian Reciprocal Ridge Regression

Consider the reciprocal bridge (rBridge) regression (Mallick et al., 2021; Alhamzawi and Mallick, 2020a; Alhamzawi, 2021) that solves the following:

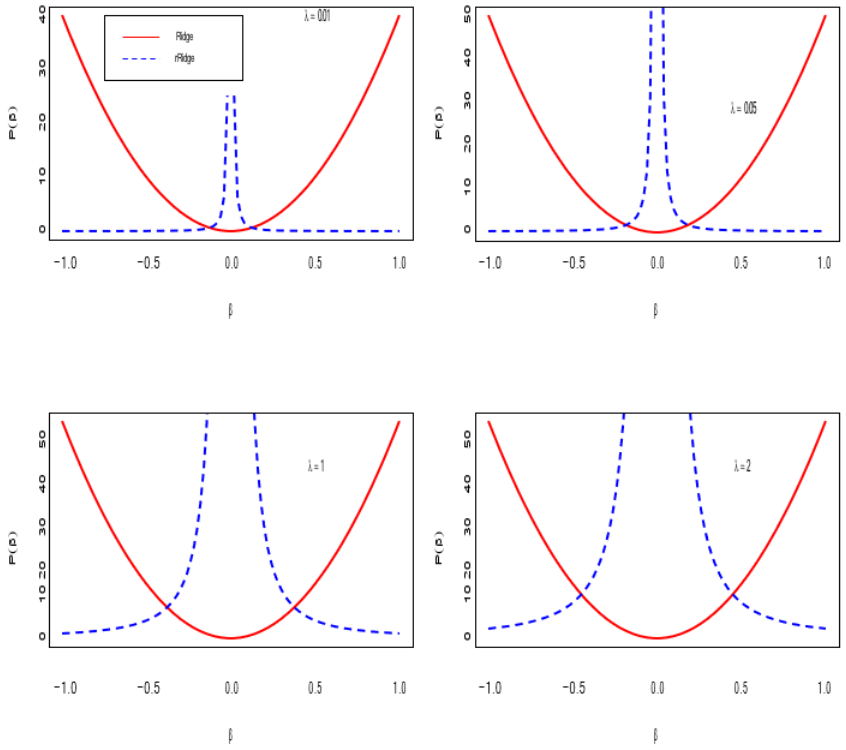

Figure 1.1: Comparison of shapes of ridge and rRidge penalty functions when

$\lambda = 0.01, 0.05, 1$ and 2, respectively.

where $I(.)$ denotes an indicator function, $\alpha > 0$ is the bridge parameter and $\lambda > 0$ is the tuning parameter that controls the degree of penalization.

Noting the form of the reciprocal penalty term in (3.2), Mallick et al. (2021) and Alhamzawi and Mallick (2020b) suggested that bridge estimates can be interpreted as posterior mode estimates when the regression parameters have independent and identical Inverse Generalized Gaussian (IGG) prior distributions of the form

$$\pi(\beta) = \frac{\lambda^{\frac{1}{\alpha}}}{2\beta^2\Gamma(\frac{1}{\alpha}+1)} \exp\{-\frac{\lambda}{|\beta|^{\alpha}}\} I\{\beta \neq 0\}, \qquad 1.20$$

where $\alpha > 0$ is a shape parameter and $\lambda > 0$ is a scale parameter. If we set $\alpha = 2$, the prior (3.3) can be used as a prior distribution for $\boldsymbol{\beta}$ corresponding to reciprocal ridge prior which can be written as:

$$\pi(\beta) = \frac{\lambda^{\frac{1}{2}}}{2\beta^2\Gamma(\frac{3}{2})} \exp\{-\frac{\lambda}{\beta^2}\} I\{\beta \neq 0\}, \qquad 1.21$$

3.1.1 MCMC Sampling

For $t = 1, \ldots, (t_{\max} + t_{\text{burn-in}})$

1. Sample $u|. \sim \prod_{j=1}^{p} \text{Exponential}(\lambda) I\{u_j > \frac{1}{|\beta_j|^2}\}$
2. Sample $\sigma^2|. \sim$ Inverse Gamma $(\frac{n-1}{2} + c, \frac{1}{2}(y^* - X\beta)'(y^* - X\beta) + d)$
3. Sample $\beta|.$ from a truncated multivariate normal proportional to

$$N_p(\hat{\beta}, \sigma^2(X'X)^{-1}) \prod_{j=1}^{p} I\{|\beta_j| > \frac{1}{u_j^{\frac{1}{2}}}\} \qquad 1.24$$

4. Sample $\lambda|. \sim \text{Gamma}(1 + p + \frac{p}{2}, \sum_{j=1}^{p} \frac{1}{|\beta_j|^2})$

Simulation studies

n	σ	ridge	BayesRidge	rBayesRidge	
150	1	1.10514	1.09679	**1.09677**	
150	2	4.12444	4.08072	**4.08064**	
150	3	9.59312	9.33292	**9.33291**	
150	4	**16.27036**	16.98162	16.98165	
150	5	**25.65290**	26.76176	26.76173	
150	6	35.61496	37.34657	**35.34648**	
200	1	1.01782	1.00373	**1.00371**	
200	2	**4.49913**	4.62405	4.62407	
200	3	9.59205	9.71667	**9.41670**	
200	4	16.95061	16.99622	**16.94617**	
200	5	27.80989	27.36470	**27.36464**	
200	6	39.53553	38.89284	**38.89280**	
250	1	**1.00738**	1.01650	1.01652	
250	2	**4.29941**	4.38438	4.38440	
250	3	9.09782	9.19361	**9.06359**	
250	4	16.08517	15.97662	**15.97658**	
250	5	**25.64371**	25.85037	25.85034	
250	6	**38.72690**	39.46833	39.46832	
300	1	1.12168	**1.03259**	1.03261	
300	2	**3.86418**	3.88761	3.88761	
300	3	**9.30678**	9.45215	9.45209	
300	4	15.43921	**15.41776**	15.71775	
300	5	24.99679	25.11152	**24.11152**	
300	6	36.90817	37.41579	**35.41580**	
500	1	0.95410	**0.95053**	0.95054	
500	2	3.86519	3.86202	**3.86201**	
500	3	**9.14054**	9.19716	9.19717	
500	4	16.36477	16.53736	**15.53316**	
500	5	25.12392	25.31819	**25.20512**	
500	6	**39.08028**	39.21197	39.21198	

This section focuses on examining the predictive accuracy of Bayesian reciprocal ridge (rBayesRidge) and comparing its performance with Bayesian ridge (BayesRidge) and frequentist ridge (ridge) in various simulated settings. The LARS algorithm developed by Efron et al. (2004) is employed for frequentist ridge regression. This algorithm utilizes 10-fold cross-validation to determine

the optimal regularization parameter (λ). The implementation of this algorithm can be found in the R package lars (Hastie and Efron, 2013). The Bayesian estimates for the ridge (BayesRidge) and reciprocal ridge (rBayesRidge) are calculated as the posterior means using 10,000 samples from the Gibbs sampler, after discarding the first 1000 samples as burn-in. The data in the simulation studies are produced by

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \sigma\varepsilon_i, \quad \text{where } i = 1, \cdots, n. \tag{1.22}$$

The outcome variable y_i is centered and the covariates are standardized to have zero means and unit variances. Methods are evaluated based on median of mean squared error (MMSE).

4.1.1 Simulation Study 1

The rows of X are simulated independently from Np(0,) with

ij = 0.95i−j. We simulate training data sets with 150, 200, 250, 300 and 500 observations and a testing data set with 200 observations. We repeat the experiment for 500 replications. The regression coefficients β and σ are setup as follows β = (1, 0, 0, 1, 1, 0, 0, 1) and σ = 1, 2, 3, 4, 5 and 6. The results of median of mean squared error (MMSE) are listed in Tabel 4.1, where the median is taken over 500 replications. The results show that, rBayesRidge performs the best. It has the smallest MMSEs in 16 out of 30 simulation setups. We also see that the ridge method performs the best in 11 out of 30 simulation setups and BayesRidge method performs the best in 3 out of 30 simulation setups. The results of the standard deviations of MSEs are shown in Figure 4.1. We are able to note that both BayesRidge and rBayesRidge demonstrate comparable effectiveness and typically the frequentist equivalent (ridge) outperform the others in terms of standard deviations of MSEs. We assess the convergence of the corresponding MCMC chain for BayesRidge and rBayesRidge by trace plots of the generated samples and calculating the Gelman-Rubin scale reduction factor (Gelman et al., 1992) using the coda package in R. We also compare both algorithm using the histograms. We compare the methods using a simulation study with n = 150 and σ = 1. From Figures 4.2, 4.3, 4.4, and 4.5, It is evident that both strategies produce comparable results.

Table 1.1: MMSEs for Simulation 4.1.1. The bold numbers correspond to the smallest MMSE in each category.

Air pollution data

In this chapter, we consider the air pollution data which is available in the truncSP package in R (Karlsson and Lindmark, 2014). A data frame with 500 observations on the following 8 variables. There are 500 observations, 7 predictor variables, and one response variable, hourly values of the logarithm of the concentration of PM10 (PM10). Predictors include The logarithm of the car count per hour (cars) and the temperature at a height of 2 meters above the ground. (temp), wind speed (meters/second) (wind.speed), the temperature difference between 25 and 2 meters above ground (degree C) (temp.diff), wind direction (degrees between 0 and 360) (wind.dir), hour of day (hour), and day number from October 1. 2001 (day). We centered the response variable and standardized the predictors. Parameter estimation of the air pollution data using different methods are listed in Table 5.1, which shows that the estimation of the regression coefficients are very closed using different methods. The results of the MSEs and the standard deviations are listed in Table 5.2, which shows that rBayesRidge outperform the others in terms of MSE and the standard deviation. We assess the convergence of the corresponding MCMC chain for BayesRidge and rBayesRidge by trace plots of the generated samples and calculating the Gelman- Rubin scale reduction factor (Gelman et al., 1992) using the coda package in R. We also compare both algorithm using the histograms. From Figures 5.1, 5.2, 5.3, and 5.4, we can see that both methods yield similar performance.

Table 1.2: Parameter estimation of the air pollution data

Variables	ridge	BayesRidge	rBayesRidge
cars	0.32505	0.35751	0.34427
temp	-0.01108	-0.01356	-0.01224
wind.speed	-0.17869	-0.19227	-0.18391
temp.diff	0.00454	0.01095	0.01077
wind.dir	-0.00362	-0.00390	-0.00871
hour	0.01781	0.00218	0.00227
day	0.04940	0.05339	0.04991

Table 1.3: MSEs and standard deviation (SD) for the air pollution data

ridge		BayesRidge	rBayesRidge
MSE	0.64708	0.64635	0.64623
SD	0.81361	0.81662	0.80143

Conclusions and FutureResearch

This thesis has compared three methods ridge, Bayesian ridge and Bayesian reciprocal ridge. Clear advantages for the existing Bayesian methods over the classical ridge include efficient MCMC-based computation techniques. Through simulation studies and analysis of an air pollution data, the performance of the Bayesian reciprocal ridgeperforms very well compared to other methods.

- In Simulation 4.1.1, the Bayesian reciprocal ridge has the smallest MMSEs in 16 out of 30 simulation setups. We also see that the ridge method performs thebest in 11 out of 30 simulation setups and BayesRidge method performs the best in 3 out of 30 simulation setups.
- In Simulation 4.1.2, the Bayesian reciprocal ridge has the smallest MMSEs in 17 out of 30 simulation setups. We also see that the ridge method performs the best in 9 out of 30 simulation setups and BayesRidge method performs the best in 4 out of 30 simulation setups.
- In Simulation 4.1.3, the Bayesian reciprocal ridge has the smallest MMSEs in 16 out of 30 simulation setups. We also see that the ridge method performs the best in 13 out of 30 simulation setups and BayesRidge method performs thebest in 1 out of 30 simulation setups.
- The air pollution data analysis show good performance for the Bayesian reciprocal ridge method.

The work considered in this thesis can be extended to compare the method discussed for to bit data, left censored data, right censored data, binary data, count data and ordinal data.

References:

1 Frank, L. E. and J. H. Friedman (1993). A statistical view of some chemometrics regression tools. *Technometrics 35* (2), 109–135.

2 Song, Q. (2018). An overview of reciprocal l_1-regularization for high dimensional regression data. *Wiley Interdisciplinary Reviews: Computational Statistics 10* (1), e1416.

3 Song, Q. and F. Liang (2015). High-dimensional variable selection with reciprocal l_1-regularization. *Journal of the American Statistical Association 110* (512), 1607–1620.

4 Hoerl, A. E., R. W. Kannard, and K. F. Baldwin (1975). Ridge regression: some simulations. *Communications in Statistics-Theory and Methods 4* (2), 105–123.

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

6 Hoerl, A. E. and R. W. Kennard (1976). Ridge regression iterative estimation of the biasing parameter. *Communications in Statistics-Theory and Methods 5* (1), 77–88.

7 Mallick, H., R. Alhamzawi, E. Paul, and V. Svetnik (2021). The reciprocal bayesian lasso. *Statistics in medicine 40* (22), 4830–4849.

8 Alhamzawi, R. and H. Mallick (2020a). Bayesian reciprocal lasso quantile regression. *Communications in Statistics-Simulation and Computation*, 1– 16.

9 Karlsson, M. and A. Lindmark (2014). truncSP: An R package for estimation of semi-parametric truncated linear regression models. *Journal of Statistical Software 57* (14), 1–19