ROBUST TECHNIQUE TO SOLVE MULTICOLINEARITY AND OUTLIER

Bahr Kadhim Mohammed and Asraa Khudair
Statistics, College of Administration and Economics, University of Al-Qadisiyah, Iraq
Emails: Bahr.mohammed@qu.edu.iq , Admin.stat21.4@qu.edu.iq

Abstract

In this paper, we modified the classical ridge regression (RR) to be more resistant for outliers in Y direction. The procedure for modification is by combining the RR with a high breakdown point and high efficiency robust methods generalized M-estimator (GM) based on robust variance covariance matrix such as MRCD. The largest advantages of the proposed method is that it has less RMSE and higher efficiency than existing methods to overcome the combined problem of multicollinearity and outliers points. A simulation study referred that the suggested method has the sprier performance around of all existing methods.

Keyword: Multicollinearity, outliers, ridge regression, robust ridge regression, M, MM and GM-estimator.

1 Introduction

Multicollinearity is a common issue in multiple linear regression, its occurs when two or more independent variables have high correlations. Multicollinearity leads to inflation of variance and destroy the coefficients of model, and give misleading conclusions [Groβ (2003)]. The other common problem is the existence of abnormal data in the dataset. Outliers also have high effects in regression model. Its accountable for model failure and misleading inference [Rousseeuw and Leroy (2003)]. in the presence of multicollinearity and outliers, the ordinary least squares estimators are unsettled and may have huge variance , which leads to misleading conclusion. The ridge regression by Hoerl and Kenard (1970) is a common approach to overcome the multicollinearity problem. Unfortunately, ridge regression is not robust to outliers. As a remedial technique, a lot of robust approaches are suggested [Huber (2003, Maronna (2006)], such as the M-estimator, the MM-estimator and the generalized M (GM-estimator). Regrettably, neither robust aproachs nor the ridge regression alone is can address the combined problem of multicollinearity and outliers [see, Habshah and Marina (2007)]. To compact this combined problem many approaches were suggested by integrating RR with some robust methods. Askin and Montgomery (1980) proposed a new technique by using weighted RR. Jadhav and Kashid (2011) suggested using ridge M-estimator to overcome multicollinearity and outliers. In this work, we propose to integrate Ridge Regression with robust method, namely GM-estimator based on high breakdown variance covariance matrix (MRCD), to overcome the multicollinearity and outliers. the matrix MRCD has A well-constructed condition, even in the ($P>n$) condition, it maintains robustness. The MRCD is a generalization of MCD when n is a number large enough to be compared to number of variables. The article is organized as follows: Section 2 presents briefly the Ridge Regression model. Section 3 explain of robust regression approach and discuss some common robust methods. Section 4 gives the structure and estimators of General M - estimate (GM). Section 5 presents the procedure of robust regression technique. Section 6 presents the simulation study with criteria's of assess the performance of methods . The discussion is presented in Section 7. Finally, Section 9 gives the conclusions.

3 Robust regression Models

A type of modern technique that aims to provide estimates that are not affected by outliers, so they often produce relatively effective estimates when the error distribution is in the normal form and it is An alternative to least squares estimators. A common robust approach is the M estimator, it's one of the resilient techniques was proposed by Huber in 1964, as it is considered one of the commonly used techniques in the linear regression model, which depends on the idea of changing the error sum of least squares (the sum of the residual function is increasing at a slower speed) instead of the squared values. These estimators reduce the impact of unusual data, as this estimator enjoys its strength in front of distant points, in contrast to its position in front of leverage points that it is sensitive to. It is an estimate that is almost equal to the efficiency of the least squares estimators, and this technique is known by another name, which is (rho function). The M estimator is given by;

$$\min_{\beta} \sum_{i=1}^{n} \rho(r_i) = \min_{\beta} \sum_{i=1}^{n} \rho\left(y_i - \sum_{i=1}^{n} x_{ij}\hat{\beta}_j\right) \qquad (1)$$

ρ represents a particular function that determines the contributions of the remainder to the target function. The other more significant approach is MM estimator. It is one of the most commonly used applications in the field of strong regression, where the high breaking point (BP=0.5) and high efficiency (95%) are combined, where the efficiency ratio refers to the efficiency of OLS under the basic assumptions.

The MM symbol indicates that more than one M estimation process is used to generate final grades, it is calculated by:

- Calculating the initial changes of the coefficients that are consistent with one of the robust methods and that have a breakdown point of (0.5) such as the estimators of S with Huber or the bisquare weight function.
- Finding the estimator M for the residuals based on the above-mentioned step.
- Using the result of the first step and the resulting scale from the second step to calculate the estimations of M-Huber based on the function Ψ.

4 General M - estimate (GM)

Humble et al (1986) explained that the M estimator does not contain VIF ,as this estimator fails to calculate the levers. To solve the problem of this failure, an estimator was proposed by Schwepp (Andersen 2008 ,Hill 1977) where this estimator produces weights in each direction X and Y ,which is an estimator known as the generalized estimator (GM), but the general form of it is known as the following form

$$\sum_{i=1}^{n} \pi_i \psi\left(\frac{y_i - x_i^t \hat{\beta}}{\hat{\sigma}\pi_i}\right) x_i \qquad (2)$$

here π denotes the initial weight function that controls the weight given to the lever .Referring to the equation, it can be solved using IRLS technology, so the form of the GM estimator becomes as follows

$$\hat{\beta}_{GM} = (X^t W X)^{-1} X^t Y \qquad (3)$$

where W is a diagonal weight matrix with finite elements defined byw_i

$$\omega_i = \frac{\psi\left[\left(y_i - x_i^{'}\hat{\beta}_{GM}\right)/\pi_i\hat{\sigma}\right]}{\left(y_i - x_i^{'}\hat{\beta}_{GM}\right)/\pi_i\hat{\sigma}} \qquad (4)$$

In 1975, Mallows developed a new strategy for generalized estimations using weights to reduce observations that have high points of influence in the form $\pi_i = \sqrt{1 - h_{ii}}$ .As for the other strategy proposed by (Karsker and Welsch) in 1975 in the form of $\pi_i = \sqrt{(1 - h_{ii})/h_{ii}}$ .Despite reaching these strategies, they are not very effective because they give low weights to good leverage points. Here, Marona et al. (1979) explained that the refractive point of the generalized estimator does not exceed) 1/p+1 ,(but despite these limitations, the GM estimator still has a high efficiency of up to 95% and close characteristics to the distribution of M estimators.

5- Robust Ridge Regression

As we explain previously, the ordinary least square has many difficulties, in order to overcome these difficulties, Hoerl and Kennard (1970) suggested an alternative technique approach called Ridge Regression (RR). This proposed approach is based on adding a bias term into the estimators to reduce their variance. They pointed that $X'X + K$, where K is a positive constant. For robust version of ridge regression, assume that robust estimated parameter is calculated by using robust approach is $\tilde{\beta}$ and for canonical by using $\alpha = \gamma^{'t}\beta$ obtain the Robust estimated parameter of canonical form $\tilde{\alpha}$. It can be said that when the Ridge method is applied, it is same as multiplied it by ($I - k\beta^{-1}$) . The Robust Ridge estimator is given by applying Ridge method to estimate parameter obtained using M , MM and GM-estimation. The estimated parameter of Robust Ridge regression for canonical form is given by:

$$\hat{\alpha}_{Robust\ ridge} = (I - k\beta^{-1})\,\tilde{\alpha} \qquad (5)$$

and the estimated parameter of Robust Ridge regression is:

$$\hat{\beta}_{Robust\ Ridge} = \gamma\,\hat{\alpha}_{Robust\ \ Ridge} \qquad (6)$$

The bias, variance and MSE of the Robust Ridge estimator are:

$$Bias(\hat{\alpha}_{Robust\ Ridge}(k)) = kB^{-1}\alpha \qquad (7)$$

$$Var(\hat{\alpha}_{Robust\ Ridge}(k)) = (I - kB^{-1})\Lambda^{-1}(I - kB^{-1})' \qquad (8)$$

$$MSE(\hat{\alpha}_{Robust\ Ridge}(k) = (I - kB^{-1})\Lambda^{-1}(I - kB^{-1})' + k^2 B^{-1}\alpha\alpha' B^{-1} \qquad (9)$$

where Λ is a variance-covariance matrix obtained by using M , MM and GM-estimation. Choosing k for Robust Ridge regression is same as when choosing k for Ridge regression. Using method that Hoerl and Kennard [1] proposed that is

$$k = k_{HK} = \frac{p\hat{\sigma}^2}{\sum_{i=1}^{p} \hat{\alpha}_i^2} \qquad (10)$$

6 Simulation study

In this section, we demonstrated a simulation experiment to compare the efficiency of the methods of study. In order to generate data with multicollinearity problem, we follow the approach of Lawrence and Arthur (1990). The multiple linear regression model is given by:

$$y_i = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \alpha_3 x_{i3} + e_i \qquad (11)$$

where e_i is the error term distributed as $N(0,\sigma^2 I)$. The dependent variables are generated as,

$$x_{ij} = \rho u_{i4} + (1-\rho^2)^{1/2} u_{ij}, \;\; i = 1,2,\cdots,n; \; j = 1,2, \text{and } 3. \qquad (12)$$

where u_{i1}, u_{i2}, u_{i3}, and u_{i4} are independent standard normal pseudo random numbers, and $k = 3$ is the number of independent variables. The ρ^2 is the degree of collinearity between $x's$. Three values of collinearity are consider ($\rho = 0.90, 0.95$ and 0.99), with four different sizes of samples (n =40, 70, 100 and 200). The contamination of data is done by replacing a clean observations by huge data in the dependent variable with different ratios of the outliers ($\tau =$ 0.01, 0.5 and 0.10). the following methods are considered- in this simulation of study

- Ridge Regression
- Robust Ridge Regression based on M-estimator (RR_Mest)
- Robust Ridge Regression based on MM-estimator (RR_MMest)
- Robust Ridge Regression based on GM- MRCD estimator (RR_GM_MRCD).

To asses the performance of the methods, the following criteria's are considered:

i) **Rote Mean Square Error (RMSE)**:

The RMSE is given as follows (Lawrence and Arthur_1990)

$$\text{RMSE}(\alpha, \hat{\alpha}) = \sqrt{E[(\hat{\alpha}-\alpha)(\hat{\alpha}-\alpha)']} \qquad (13)$$

$$\text{RMSE}(\alpha_j) \;\; = \sqrt{\frac{1}{R}\sum_{i=1}^{R}(\hat{\alpha}_{ij}-\alpha_j)^2}, \; j = 1,2,...,k \qquad (14)$$

where, R = 5000 is a replication of Mont simulation experiments, $\hat{\alpha}_{ij}$ is the ith estimate of the jth parameter in the ith replication, and α_j, j = 1, 2, and 3, are the true coefficients of the regression model chosen as $\alpha_1 = 1, \; \alpha_2 = 1.5, \; \textit{and} \; \alpha_3 = 1.7$

ii) **Efficiency**

Comparison of The RMSE ratios [Jadhav and Kashid (2013)] of our proposed method (RR_GM.MRCD) over of Ridge, RR_Mest, and RR_MMest are calculated for all possible combinations of n, ρ and τ . If the ratio is less than one, the our proposed method is more efficient than the other method.

Result and discussion

Tables 1 -3 present the results of RMSE and Efficiency of simulation study.

Table 1: RMSE and Relative efficiency for estimation methods with contamination 1%

Methods	N	40		70		100		200	
	ρ	RMSE	Eff.	RMSE	Eff.	RMSE	Eff.	RMSE	Eff.
Ridge	0.90	0.133	0.512	0.124	0.510	0.121	0.511	0.121	0.490
RR_Mest.		0.131	0.514	0.125	0.521	0.122	0.513	0.121	0.489
RR_MMest.		0.112	0.603	0115	0.554	0.116	0.533	0.115	0.516
RR_GM_mrcd		0.106	1	0.110	1	0.112	1	0.111	1
Ridge	0.95	0.136	0.547	0.126	0.552	0.123	0.553	0.122	0.524
RRidge_Mest.		0.132	0.564	0.125	0.554	0.123	0.553	0.122	0.524
RRidge_MMest.		0.113	0.579	0.107	0.598	0.118	0.577	0.116	0.568
RR_GM_mrcd		0.108	1	0.102	1	0.113	1	0.113	1
Ridge	0.00	0.161	0.585	0.135	0.637	0.129	0.649	0.126	0.591
RRidge_Mest.		0.137	0.685	0.127	0.674	0.125	0.669	0.124	0.599
RRidge_MMest.		0.123	0833	0.117	0.758	0.118	0.710	0.117	0.634
RR_GM_mrcd		0.111	1	0.103	1	0.115	1	0.114	1

Table 2: RMSE and Relative efficiency for estimation methods with contamination 5%

Methods	N	40		70		100		200	
	ρ	RMSE	Eff.	RMSE	Eff.	RMSE	Eff.	RMSE	Eff.
Ridge	0.90	0.143	0.471	0.135	0.468	0.136	0.493	0.135	0.439
RR_Mest.		0.140	0.483	0.131	0.471	0.135	0.501	0.135	0.439
RR_MMest.		0.103	0.653	0.105	0.604	0.102	0.653	0.101	0.589
RR_GM_mrcd		0.099	1	0.100	1	0.098	1	0.098	1
Ridge	0.95	0.149	0.499	0.138	0.504	0.134	0.493	0.130	0.471
RRidge_Mest.		0.140	0.530	0.135	0.516	0.132	0.501	0.126	0.477
RRidge_MMest.		0.105	0.710	0.107	0.684	0.104	0.653	0.103	0.624
RR_GM_mrcd		0.101	1	0.103	1	0.101	1	0.096	1
Ridge	0.99	0.192	0.490	0.158	0.545	0.153	0.549	0.145	0.512
RRidge_Mest.		0.151	0.621	0.138	0.623	0.138	0.605	0.137	0.544
RRidge_MMest.		0.105	0.892	0.101	0.802	0.105	0.799	0.104	0.715
RR_GM_mrcd		0.095	1	0.091	1	0.088	1	0.083	1

Table 3: RMSE and Relative Efficiency for estimation methods with contamination 10%

Methods	N	40		70		100		200	
	ρ	RMSE	Eff.	RMSE	Eff.	RMSE	Eff.	RMSE	Eff.
Ridge	0.90	0.161	0.421	0.149	0.424	0.145	0.424	0.145	0.410
RR_Mest.		0.151	0.448	0.145	0.437	0.142	0.433	0.140	0.413
RR_MMest.		0.109	0.621	0.096	0.661	0.094	0.783	0.088	0.699
RR_GM_mrcd		0.088	1	0.085	1	0.080	1	0.077	1
Ridge	0.95	0.174	0.428	0.155	0.449	0.150	0.453	0.147	0.437
RRidge_Mest.		0.154	0.484	0.146	0.478	0.144	0.471	0.138	0.447
RRidge_MMest.		0.105	0.711	0.095	0.733	0.093	0.727	0.090	0.711
RR_GM_mrcd		0.090	1	0.084	1	0.081	1	0.076	1
Ridge	0.99	0.256	0.367	0.198	0.435	0.180	0.466	0.165	0.451
RRidge_Mest.		0.182	0.515	0.155	0.555	0.151	0.556	0.147	0.504
RRidge_MMest.		0.106	0.887	0.096	0.898	0.093	0.899	0.090	0.811
RR_GM_mrcd		0.097	1	0.092	1	0.087	1	0.080	1

From the results of above tables, with diferent ratios of contaminations (1%, 5% and 10%), we can see that the (RR_GM_mrcd) has best performance above all of other methods through it has less values of RMSE and high relative efficiency followed by (RR_MMest.) . Whereas the Ridge regression method has the lowest performance. In addition we clearly notice that the values of RMSE are increased when the collinearity are increased for all of methods and all size of samples. On the other hand, it is surprising to see that all the traditional methods decrease their performance and relative efficiency with an increase in the contaminations rate, except for the proposed method, where its performance is increase with an increase of contaminations. Figures 1-6 confirm the above results.

9. Concussion

In this study, we proposed a new estimation methods called RR_GM.MRCD to remedy the combined problem of multicollinearity and outliers. The procedure for modification is by combining the RR with a high breakdown point and high efficiency robust methods generalized M-estimator (GM) based on robust variance covariance matrix such as MRCD. In order to assess the performance of the proposed method, we compared it with existing methods by using a simulation data based on RMSE and relative efficiency. The results indicate that the suggested method has superior performance compared with other methods for all cases of size of samples, collinearity ratios and contaminated Percentage.

References

[1] Hoerl E. and R. Kennard W. (1970), “Biased Estimation for Non-orthogonal Problems”, Technometrics, 12, 69-82

[2] Quenouille M. H. (1970), “Trust Notes on Bias in Estimation”, Biometrika, 53, 353-360.

[3] McDonald G. C. and Galarneau D. I. (1975), “A Monte Carlo evaluation of some ridge-type estimators”, JASA, 20, 407-416.

[4] Hinkley D. V.(1977) ,“Jackknifing in Unbalanced Situations”, Technometrics, 19, 285-292.

[5] Askin R. G. and Montgomery D. C. (1980), “Augmented Robust Estimators, Technometrics”, 22, 333-341

[6] Singh B., Chaubey Y. P. and Dwivedi T. D. (1986), “An almost unbiased ridge estimator”, The Indian Journal of Statistics, 48, 342-346

[7] Lawrence K. D. and Arthur J. L. (1990), “Robust Regression: Analysis and Applications”, Marcel Dekker, New York

[8] Huber, P.J (2003), “Robust Statistics”, Wiley, New York, USA.

[9] Maronna R. (2003), “Robust Statistics”, Wiley, New York, USA.

[10] Rousseeuw P.J. and Leory A.M. (2003),” Regression and Outlier Detection”, Wiley, New York, USA.

[11] Groβ J. (2003), “Linear Regression- Lecturer Notes in Statistics”, Springer Verlag Berlin Heidelberg.

[12] Habshah M. and Marina Z. (2007), “A Simulation study on ridge regression estimation in the presence of outliers and multicollinearity”, Journal Teknologi, 47, 59-74.

[13] Batah F. S., Ramanathan T. V. and Gore S. D. (2008), “The efficiency of modified jackknife and ridge type regression estimators: a comparison”, Surv Math Appl, 3 (2008) 111-122

[14] Jadhav N. H. and Kashid D. N. (2011), “A Jackknifed Ridge M-Estimator for Regression Model with Multicollinearity and Outliers”, Journal of Statistical Theory and Practice, 5, 659-673.

[15] Hekimoglu S. and Erenoglu R.C. (2013), “A new GM-estimate with high breakdown point”, Acta Geod Geophys, 48, 419- 437.

Figure 1: Relative Efficiency of methods with 1% of contamination

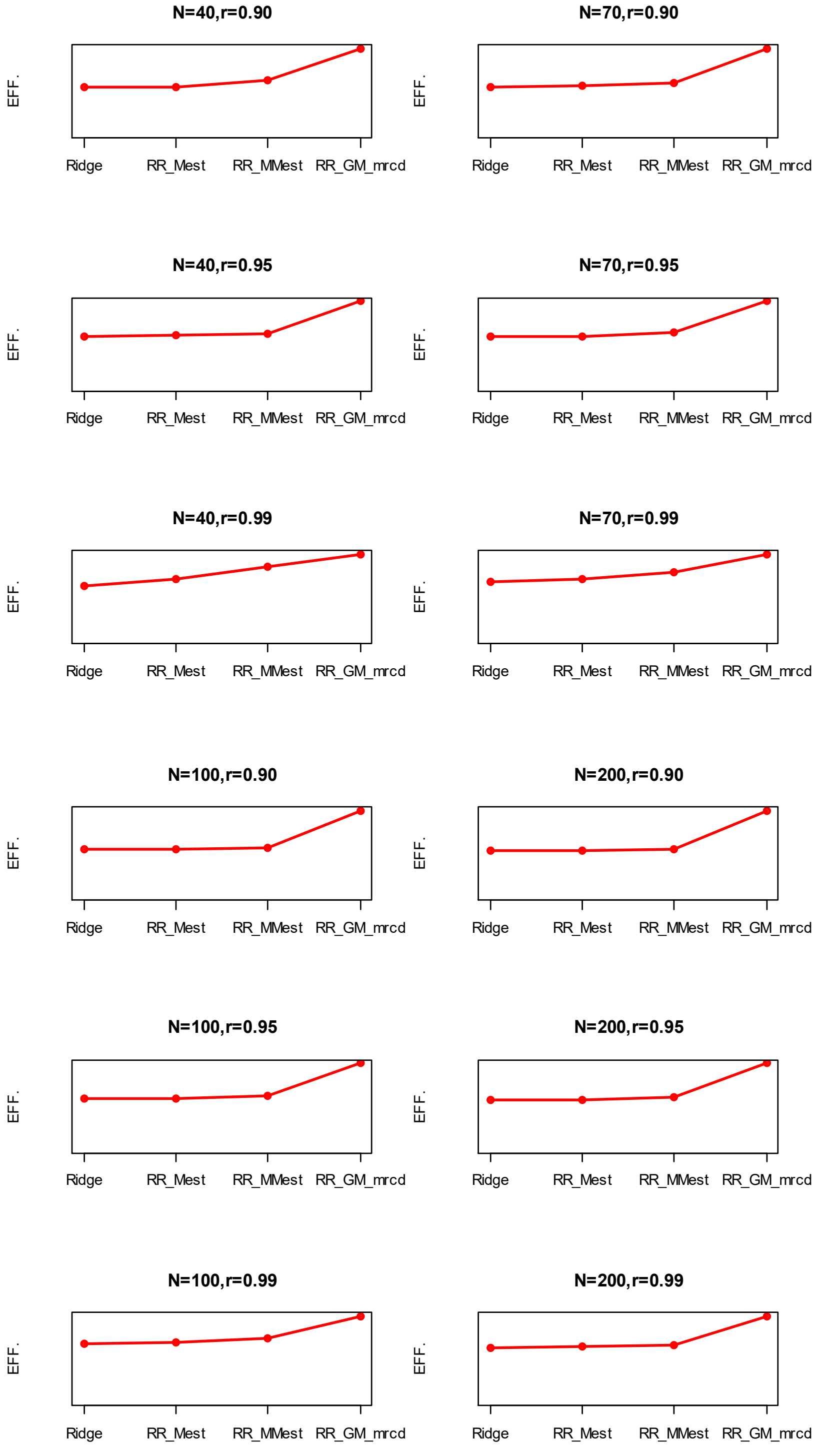


Figure 2: RMSE values of methods with 1% of contamination

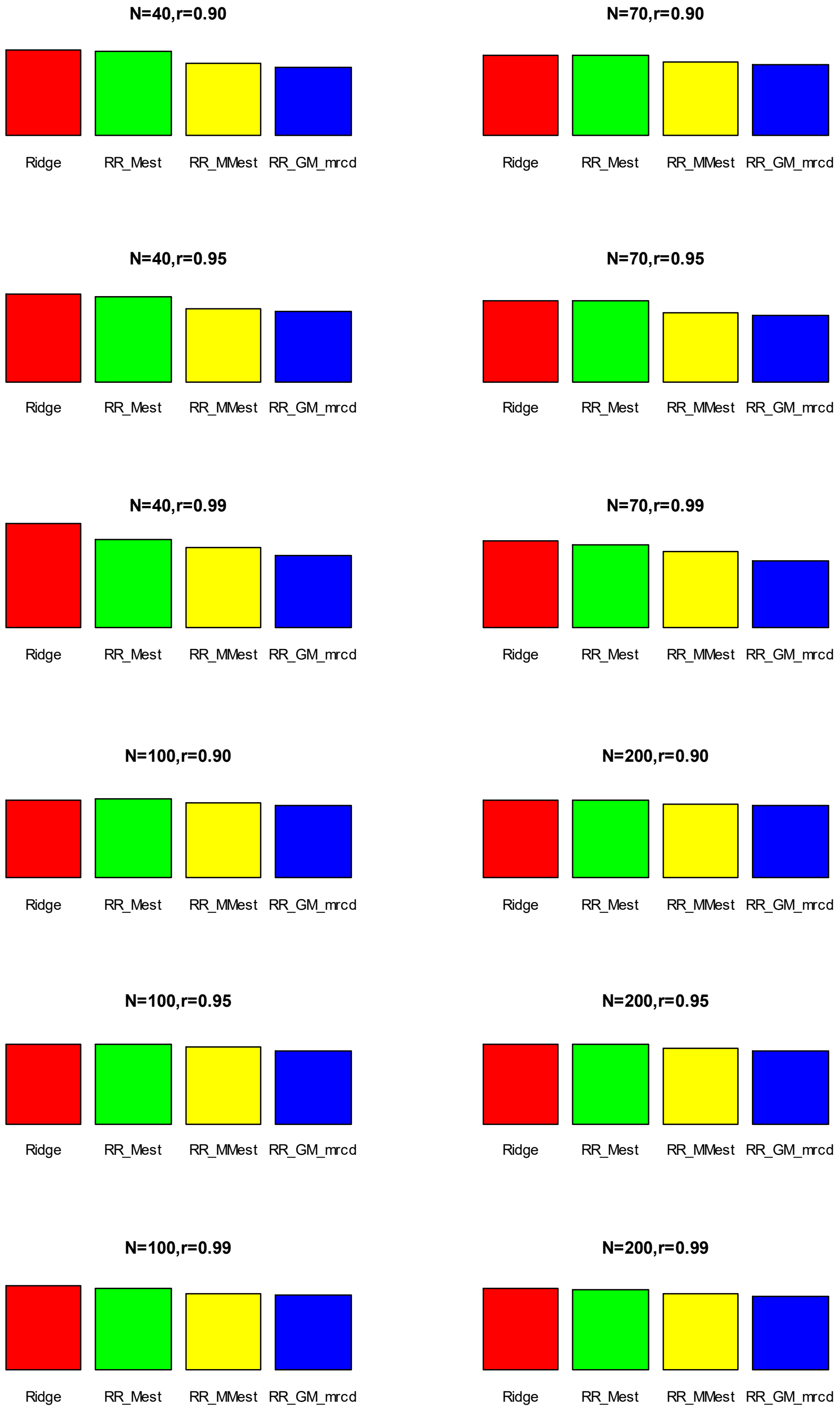
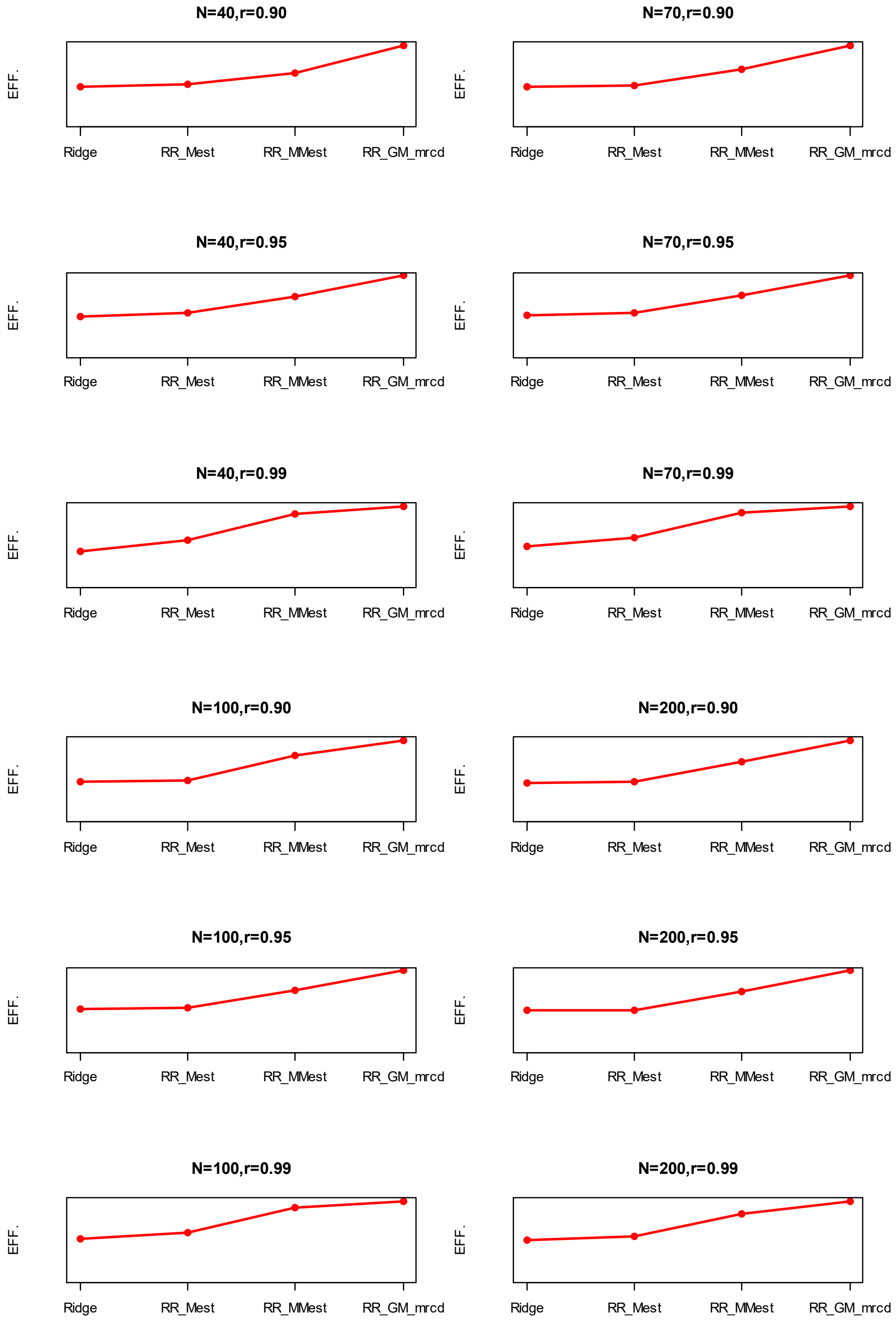


Figure 3: Relative Efficiency of methods with 5% of contamination



N=40,r=0.90

N=70,r=0.90

N=40,r=0.95

N=70,r=0.95

N=40,r=0.99

N=70,r=0.99

N=100,r=0.90

N=200,r=0.90

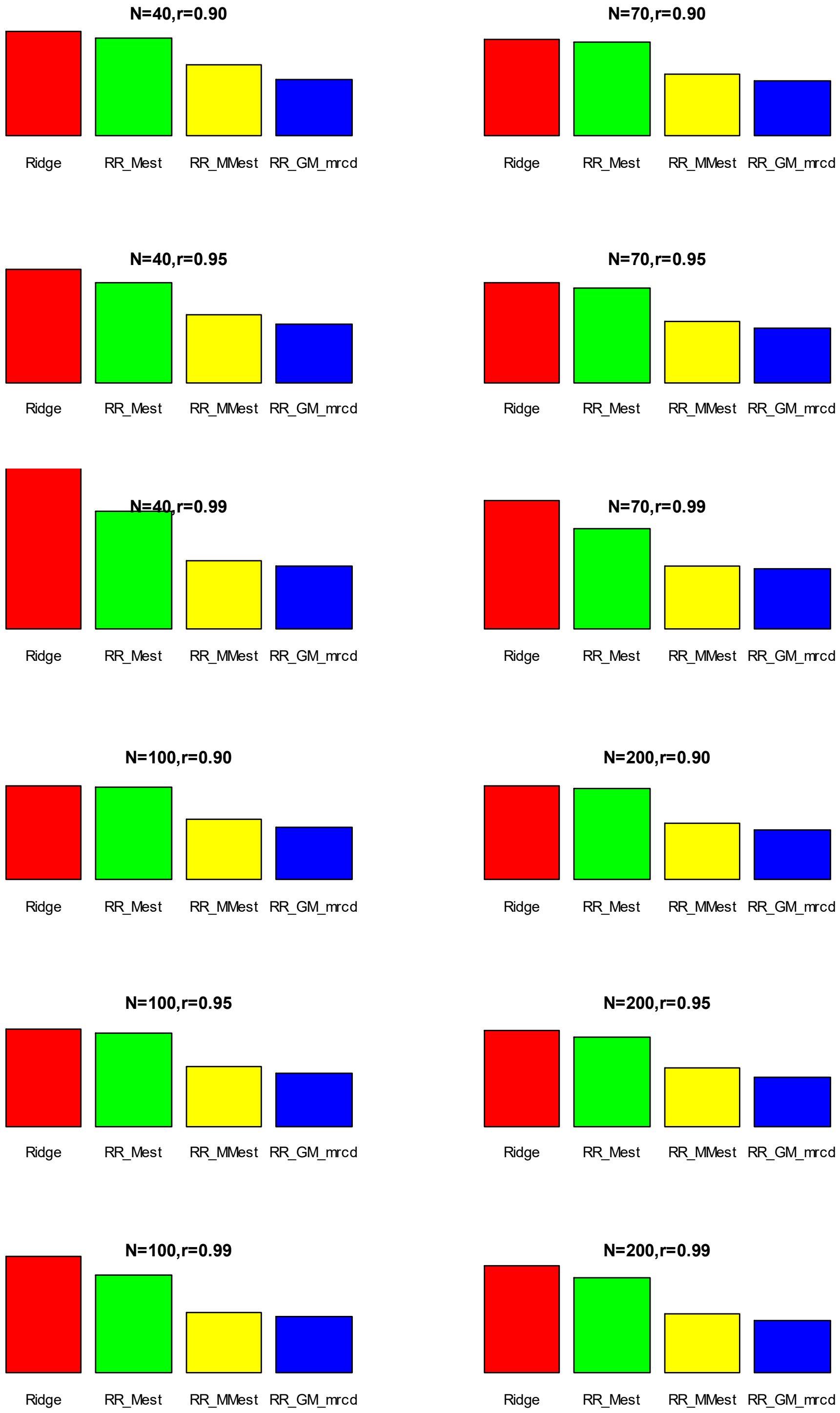
N=100,r=0.95

N=200,r=0.95

N=100,r=0.99

N=200,r=0.99

Figure 4: RMSE values of methods with 5% of contamination



N=40,r=0.90

N=70,r=0.90

N=40,r=0.95

N=70,r=0.95

N=40,r=0.99

N=70,r=0.99

N=100,r=0.90

N=200,r=0.90

N=100,r=0.95

N=200,r=0.95

N=100,r=0.99

N=200,r=0.99

Figure 5: Relative Efficiency of methods with 10% of contamination

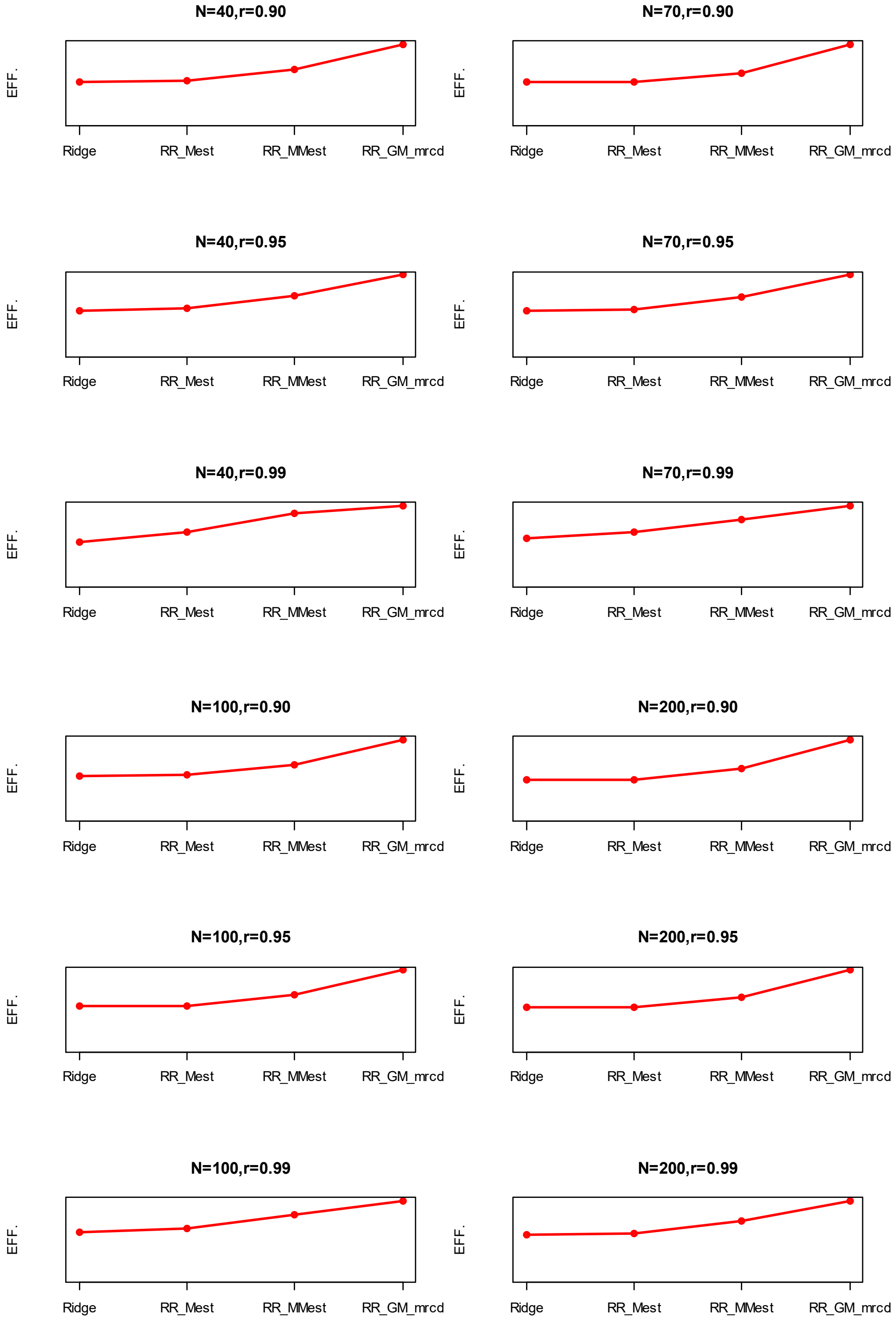
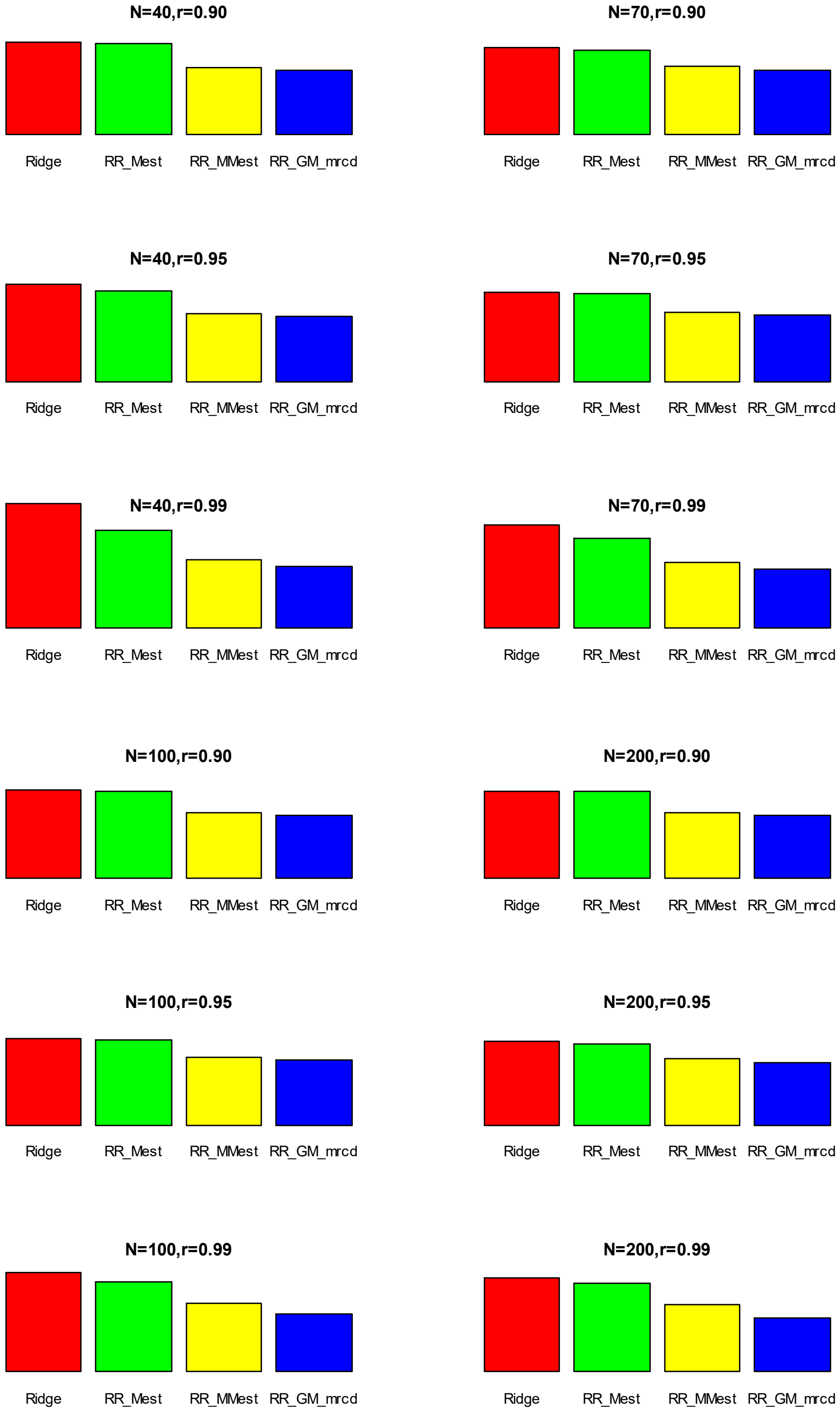


Figure 6: RMSE values of methods with 10% of contamination



N=40,r=0.90

N=70,r=0.90

N=40,r=0.95

N=70,r=0.95

N=40,r=0.99

N=70,r=0.99

N=100,r=0.90

N=200,r=0.90

N=100,r=0.95

N=200,r=0.95

N=100,r=0.99

N=200,r=0.99