NNRMLR: A Combined Method of Nearest Neighbor Regression and Multiple Linear Regression

Hideo Hirose, Yusuke Soejima
School of Computer Science and Systems Engineering
Kyushu Institute of Technology
Fukuoka, 820-8502 Japan
Email: hirose@ces.kyutech.ac.jp

Kei Hirose
School of Engineering Science
Osaka University
Toyonaka, 560-0043 Japan
Email: hirose@sigmath.es.osaka-u.ac.jp

***Abstract*—To predict the continuous value of target variable using the values of explanation variables, we often use multiple linear regression methods, and many applications have been successfully reported. However, in some data cases, multiple linear regression methods may not work because of strong local dependency of target variable to explanation variables. In such cases, the use of the k nearest-neighbor method (k-NN) in regression can be an alternative. Although a simple k-NN method improves the prediction accuracy, a newly proposed method, a combined method of k-NN regression and the multiple linear regression methods (NNRMLR), is found to show prediction accuracy improvement. The NNRMLR is essentially a nearest-neighbor method assisted with the multiple linear regression for evaluating the distances. As a typical useful example, we have shown that the prediction accuracy of the prices for auctions of used cars is drastically improved.**

***Keywords*-Linear regression; ridge; lasso; elastic net; nearest neighbor regression; combined method of linear regression and k-NN; auction price.**

I. Introduction

To predict the value of target variable y using the values of explanation variables X, we often use multiple linear regression methods (MLR), $y = X\beta$, and many applications have been successfully reported. Here, β expresses the parameters [4]. To assess the prediction accuracy of y to new X, we usually perform the two procedures: constructing the prediction formula (structure) using the training data, and assessing the accuracy by applying the prediction formula to the test data. The cross-validation method or the bootstrap out-of-sample method [5] is often used to evaluate the accuracy statistically. To improve the accuracy for test data, a variety of regularization methods in which penalty functions are imposed are proposed [4]; e.g., the ridge, lasso, elastic net, and adaptive lasso are among them.

However, in some data cases, MLR may not work because of strong local dependency of target variable to explanation variables. For example, in predicting the prices for auctions of used cars, even the use of the regularization methods such as the ridge, lasso, and their relatives, could not improve the prediction accuracy much [6]. Figure 1 shows a typical case of used car auction expressing the correlations among some selected feature variables X and target function y. This is a case of one type (model) of cars, where the sample size N is about 4,000, and the number of feature variable is more than 200 in the original data. The explanation variables x_j are mileage, color, equipments, etc. However, the concrete explanation variable names and their values are not shown for simplicity (only “ex.v.1 ... ” are shown); the numbers in boxes are the correlation coefficients. The reason why we have dealt with the car auction problem to each car model is that we have experienced that dealing with many types of cars all together at once leads us in a mess.

In such applications, the use of the k nearest-neighbor method (k-NN) [1] in regression can be an alternative. In collecting k nearest-neighbor cars, the more similar the cars, the more accurate in prediction. Even a simple k-NN regression method could improve the prediction accuracy to some extent, which indicates promising methods in this way. However, we have found that k-NN regression methods assisted with multiple linear regressions for evaluating the distances show improvements for prediction accuracy. Therefore, we propose a newly developed combined method of k-NN regression and multiple linear regression methods (NNRMLR) in this paper. Using the proposed method, the prediction accuracy is improved drastically in a used car auction example case.

In this paper, we have also investigated the prediction accuracies in the relation among the ratio of the size of training data to that of the test data (we abbreviate the latter as RTT); the detail of the selection for training and test data is explained later.

The remainder of the paper is organized as follows. In the next section, the multiple linear regression methods with regularization methods are briefly reviewed; section 3 explains the k-NN regression method; section 4 deals with a newly proposed method, a combined method of k-NN regression and multiple linear regression methods (NNRMLR); and section 5 summarizes the conclusions of this study.

978-0-7695-4826-5/12 $26.00 © 2012 IEEE
DOI 10.1109/IIAI-AAI.2012.76

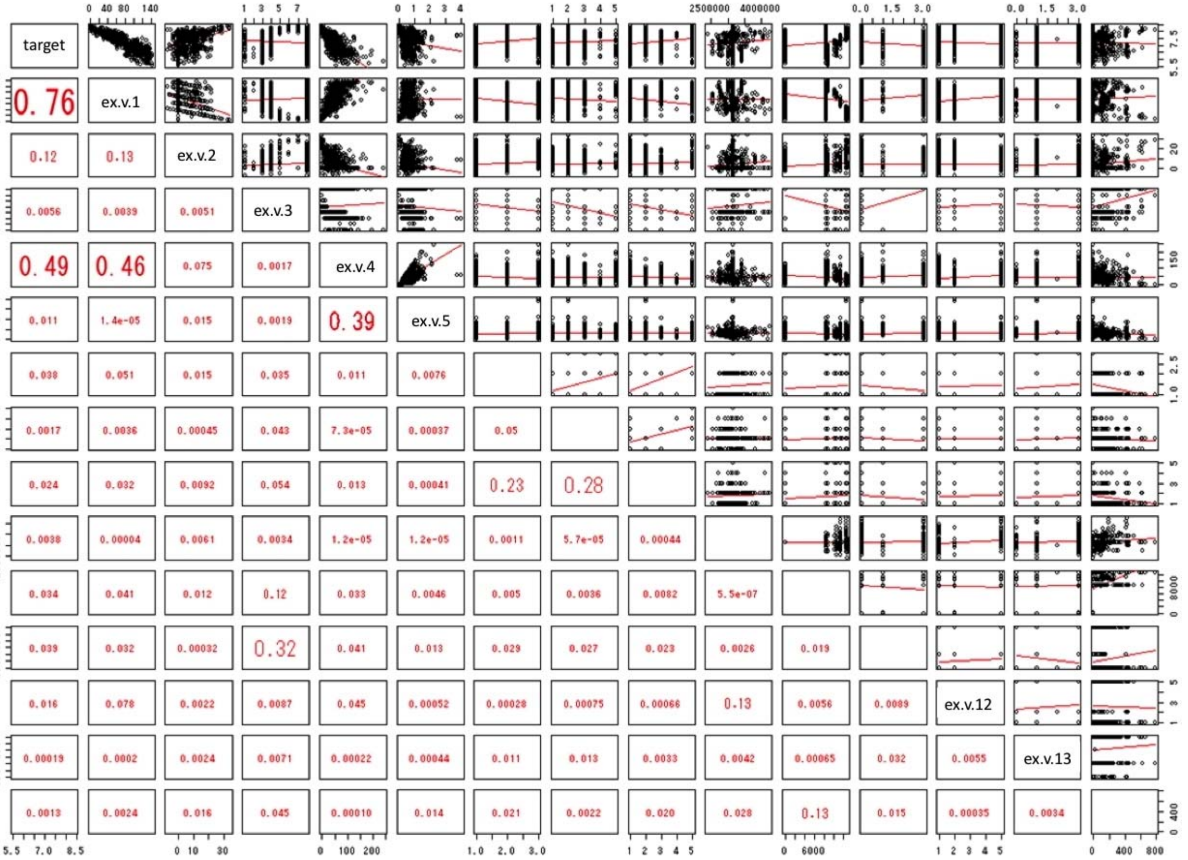

Figure 1. Correlations among feature variables and the target function.

II. Multiple Linear Regression Methods, MLR

As a typical example case, we deal with the prediction of the prices for auctions of used cars. In such cases, we often use linear regression methods, $Y = X\beta$, where the objective variable y is the price and the explanation variables x_j are milage, color, equipments, and etc. In the example case shown in Figure 1, although the number of feature variable exceeds 200, we have reduced the number of variables to 28 in advance by pre-assessment.

To find the accurate estimates for prices, we often apply regularization methods in linear regressions such as the ridge, lasso, and their relatives [4]. The optimal parameters $\hat{\beta}$ can be obtained for each regularization penalty by,

$$\hat{\beta}_{\mathrm{norm}} = \arg\min_{\beta}\{\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \mathrm{p}_{\mathrm{norm}}\}. \tag{1}$$

Here, p is the number of feature variables, and $\mathrm{p}_{\mathrm{norm}}$ (penalty term) expresses a regularization factor which changes its style according to the norm such as,

$$\mathrm{p}_{\mathrm{lasso}} = \lambda \textstyle\sum_{j=1}^{p} |\beta_j|, \tag{2}$$

$$\mathrm{p}_{\mathrm{ridge}} = \lambda \textstyle\sum_{j=1}^{p} \beta_j^2, \tag{3}$$

$$\mathrm{p}_{\mathrm{elastic\ net}} = \lambda \textstyle\sum_{j=1}^{p} (\alpha\beta_j^2 + (1-\alpha)|\beta_j|), \tag{4}$$

$$\mathrm{p}_{\mathrm{adaptive\ lasso}} = \lambda \textstyle\sum_{j=1}^{p} w_j|\beta_j|, \tag{5}$$

where, $w = 1/|\beta|^{\gamma}$, and $\lambda \geq 0$, $0 \leq \alpha \leq 1$, and $\gamma \geq 0$ are tuning parameters.

We first construct the prediction formula (structure) using the training data, and then assess the prediction accuracy by applying the prediction formula to the test data. The training data are selected randomly but different from each other, and the test data are collected from the rest of the sampled data. The ratio of the size of training data to that of the test data (RTT) are 1:9, 2:8, 3:7, 4:6, 5:5, 6:4, 7:3, 8:2, 9:1. The resampling is repeated 20 times. The training data are applied to equation (1).

The prediction accuracy is assessed by the RMSE (root mean squared error) for prediction error using the test data, which is expressed as

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}, \tag{6}$$

where, $\hat{y}_i$ and y_i are the predicted and observed values. Here, we have investigated, 1) simple MLR, 2) stepwise method [2], 3) ridge [3], 4) lasso [7], and 5) elastic net [9].

Figure 2 shows the RMSE for each method via the box-plot expression. We might expect that the larger the size of training data, the smaller the value of RMSE for test data [4]. However, too small size of test data may lose the accuracy. From the figure, we can see that best accuracy can be obtained when the RTT are around 3:7 - 7:3. In addition, we see that any regularization methods could not improve the accuracy much comparing to the simple MLR. The regularization is incompetent here. The best performance is obtained when we use the elastic net. The RMSE value is around 345.

III. Nearest Neighbor Regressions

Figure 1 indicates us to use the k-NN regression because a strong local dependency of target variable to explanation variables is seen. Some expertise in this fields also suggests this tendency. Thus, we use the k-NN regression methods first. Here, the k-NN regression means that we use mean value of the target values of the nearest feature variable points to the target prediction point in a Euclid sense such as

$$d_0(a_i, b_j) = \sqrt{\sum_{l=1}^{p} (a_{il} - b_{jl})^2}, \tag{7}$$

where, a_i and b_j denote the training and test points.

Figure 3 shows the RMSE using the simple k-NN methods; the number of nearest samples, k, are set to 1, 3, 5, 10, 50, 100. As easily expected, the RMSE is likely to be reduced as the number of nearest samples becomes smaller. Except the case when the RTT is 1:9, the best performance is obtained when $k = 1$. However, taking into account of robustness (low variance as well), we adopt the case of $k = 3$ here. As the RTT moves from 1:9 to 9:1, the RMSE becomes smaller. Here, taking into account the common use of RTT value, we adopt the cases of RTT = 5:5 and 7:3 (this is close to 3:1). Then RMSE are 250 - 300, which shows dramatically reduced values comparing to the MLR with regularizations. We can see that the k-NN regression methods are promising as expected.

IV. NNRMLR: Combined method of k-NN regression and multiple linear regression

We have just experienced the improvements of the RMSE dramatically by adopting the k-NN. However, we can further pursue the improvement by combining the MLR with regularization methods and the k-NN. We may select the more effective points in prediction; standardized $\hat{\beta}$ obtained in the MLR indicates the importance in selection of the k-NN points. We may impose the weighting in adopting the k-NN points.

The method is as follows: We first perform the MLR with some regularization method, and obtain the effective feature variables (as a weighting function). Then, using these values of β_l, we redefine the weighted distance $d_1(a_i, b_j)$ between the two points as

$$d_1(a_i, b_j) = \sqrt{\sum_{l=1}^{p} |\hat{\beta}_l| (a_{il} - b_{jl})^2}, \tag{8}$$

where $\hat{\beta}$ is the estimated value obtained by regularization methods. This means that the important variables are actively adopted in searching for the k-NN points, and the vanished feature variables due to regularizations such as the lasso are eventually not used. The results using this combination methods will be shown later.

We may further impose an optimal weight to each feature variable point, such as

$$d_\gamma(a_i, b_j) = \sqrt{\sum_{l=1}^{p} |\hat{\beta}|^\gamma (a_{il} - b_{jl})^2}, \tag{9}$$

where $\gamma \geq 0$ could be an acceleration coefficient for weighting. To find the optimal value of γ, we have obtained the RMSE for $\gamma = 1, 1.5, 2, 2.5, 3, 3.5, 4, 5$ as shown in Figure 4. The figure reveals that $2 \leq \gamma \leq 3$ can be optimal for all the MLR methods.

Finally, we show the results of proposed method, the NNRMLR, in Figure 5. In the figure, we can see the special case of "RTT 7:3" and "simple 3-NN($\gamma = 0$)", which is expressed by equation (8). The RMSE incorporating the optimal γ are shown in the figure with optimal values of γ. Among these, the k-NN with optimal weighting via the elastic net was found to be the best in our experiment. Lastly, we mention that although the RMSE of the adaptive lasso [8] (not presented here) is not much reduced comparing to other methods, the number of feature variables was reduced to almost $\frac{1}{3}$ of the initial condition.

V. Concluding remark

To predict the continuous value of target variable using the values of explanation variables, we often use multiple linear regression methods, and many applications have been successfully reported. However, in some data cases, multiple linear regression methods will not work because of strong local dependency of target variable to explanation variables. For example, in predicting the prices for auctions of used cars, even the use of the regularization methods such as the ridge, lasso, and their relatives, could not improve the prediction accuracy much.

In such applications, the use of the k nearest-neighbor method in regression is an alternative. Although a simple k nearest-neighbor method improves the prediction accuracy, a newly proposed method, a combined method of k nearest-neighbor regression and linear regression methods

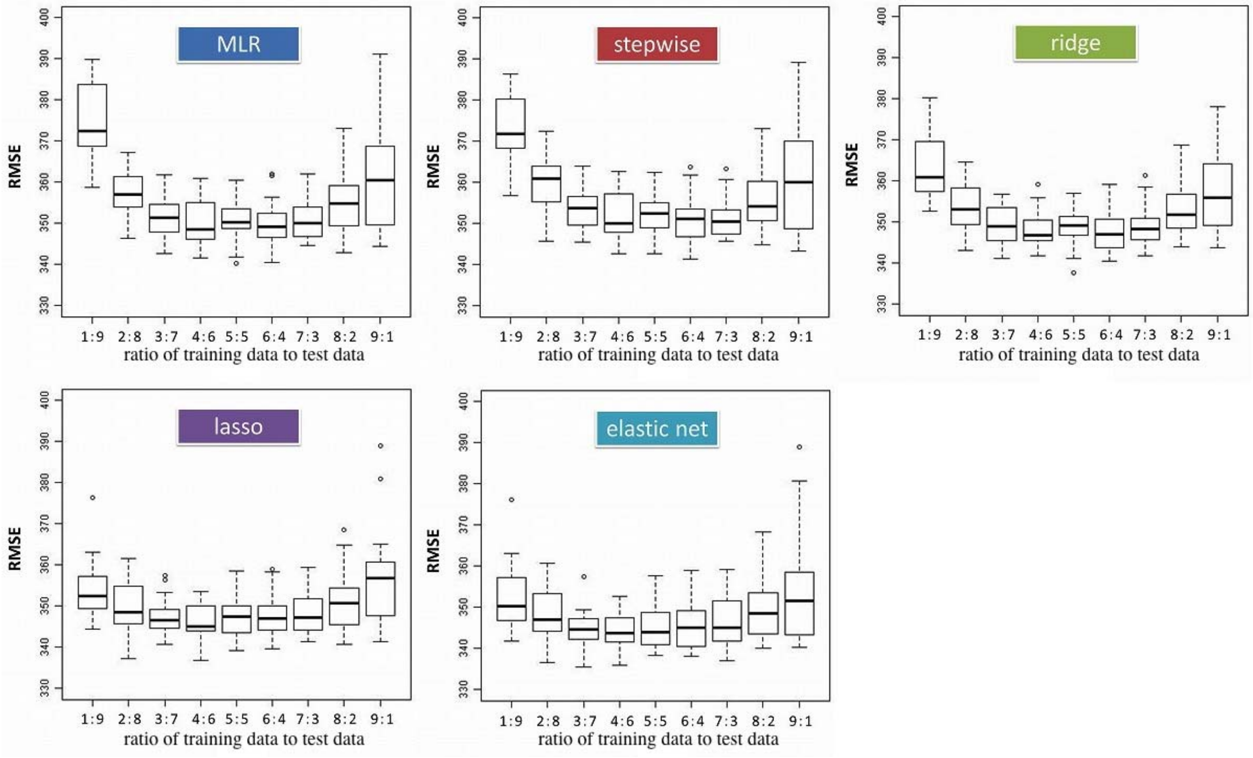

Figure 2. RMSE by the MLR with various regularization methods.

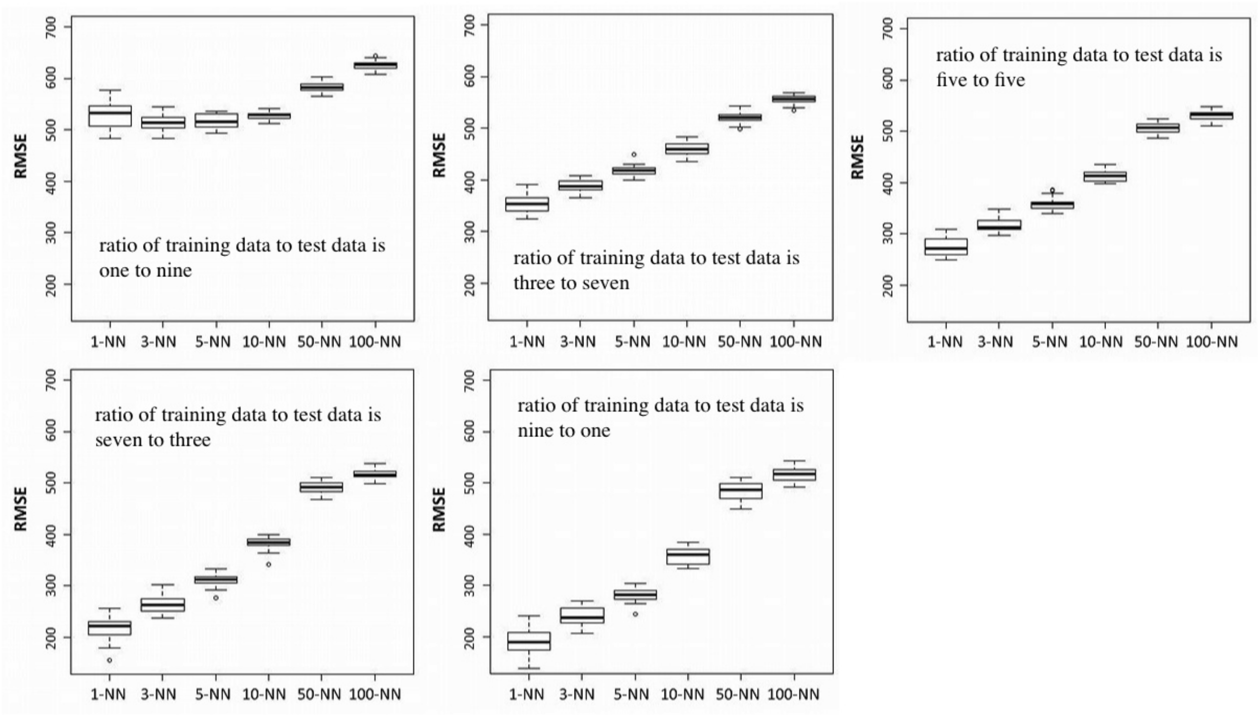

Figure 3. RMSE by the simple k-NN when the ratio of the size of training data to that of the test data are 1:9, 3:7, 5:5, 7:3, 9:1.

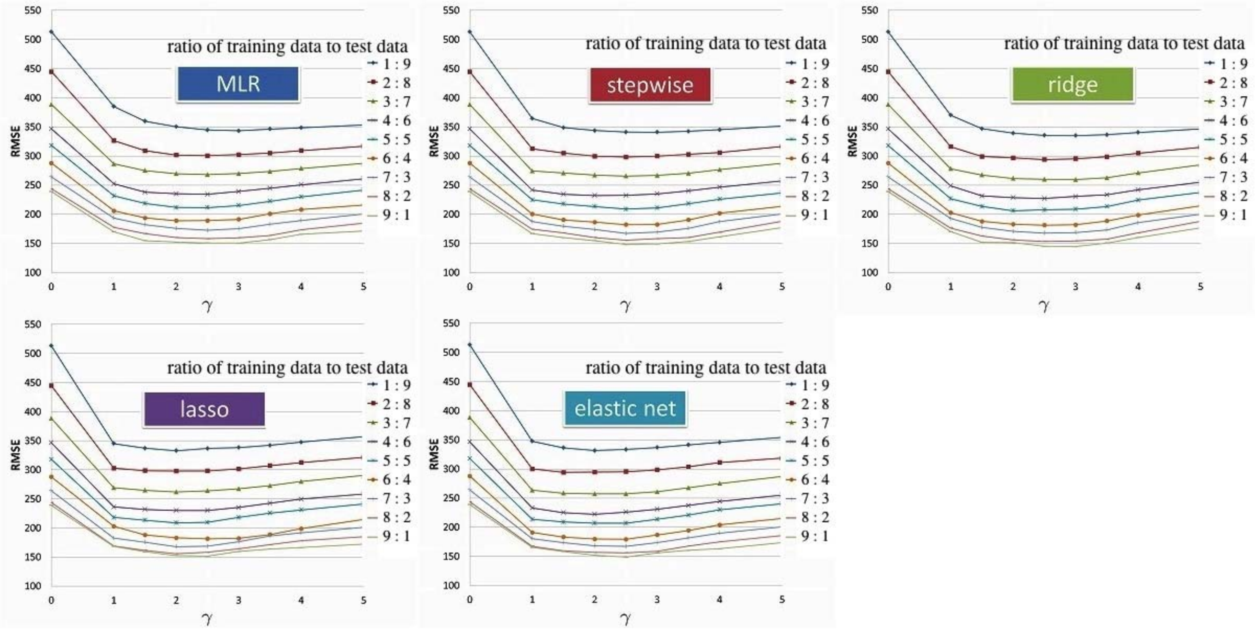

Figure 4. RMSE by MLR with regularization versus weighting value γ.

(NNRMLR), is found to show prediction accuracy improvements. That is, we first perform the multiple linear regression with some regularization method, and obtain the effective feature variables (as a weighting function). Then, using the values of estimated parameters, we redefine the weighted distance between the two points. We have also investigated the relation between the accuracy and the ratio of the size of training data to that of the test data. The commonly used ratios of 5:5 or 7:3 are used here.

The optimum accuracy could be obtained stably when 1) the ratio of the size of training data to that of the test data is 7:3, 2) the number of nearest neighbors is 3, 3) weighting parameter for distance function is around 2-3. Then, the RMSE obtained by the multiple linear regression methods with the regularizations, around 345, can be reduced to 260. Moreover, incorporating the weighting parameter for distance function, the RMSE is reduced to less than 200. This is a dramatic improvement.

References

[1] Dasarathy, B. V., *Nearest Neighbor (NN) Norms; NN Pattern Classification Techniques*, IEEE Computer Society (1991).

[2] Efroymson, M., Multiple regression analysis, *Mathematical Methods for Digital Computers*, Vol. 1, pp. 191–203 (1960).

[3] Hoerl, A. E. and Kennard, R. W., Ridge Regression: Biased Estimation for Nonorthogonal Problems, *University of Delaware and E. I. du Pont de Nemours* & *Co*, Vol. 12, No. 1 (1970).

[4] Hastie, T., Tibshirani. R., and Friedman, J., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, 2 edition (2009).

[5] Hothorn, T., Leisch, F., Zeileis, A., Hornik, K. The Design and Analysis of Benchmark Experiments, Journal of Computational and Graphical Statistics, 14, pp.675-699, (2005).

[6] Soejima, Y., and Hirose, H., Auction price estimation for used cars by regression methods, Joint Meeting of the Korea-Japan Conference of Computational Statistics and the 25th Symposium of Japanese Society of Computational Statistics, pp.9-12, 2011.

[7] Tibshirani, R., Regression shrinkage and selection via the lasso, *the Royal Statistical Society*, Vol. 58, pp. 267–288 (1996).

[8] Zhang, H. and Lu, W., Adaptive lasso for Cox's proportional hazards model, *Biometrika*, Vol. 94, pp. 691–703 (2011).

[9] Zou, H. and Hastie, T., Regularization and variable selection via the elastic net, *Journal of the Royal Statistical: Series B*, No. 67(2), pp. 301–320 (2005).

[10] Zou, H., The adaptive lasso and its oracle properties, *JASA*, Vol. 101, No. 476, pp. 1418–1429 (2006).

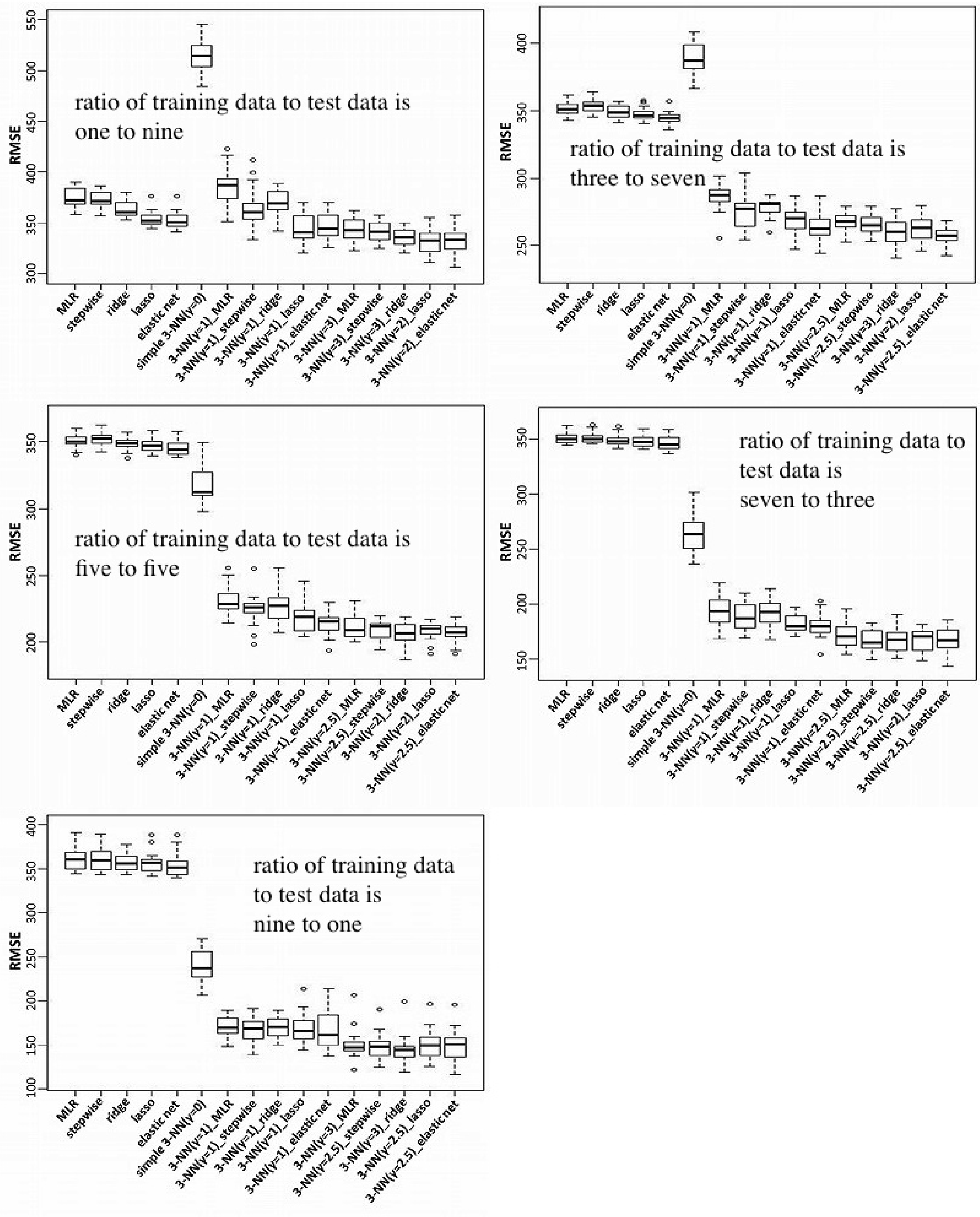

Figure 5. RMSE comparison by all the methods.