J. Appl. Math. & Informatics Vol. **27***(2009), No. 3 - 4, pp. 903 - 908*
Website: http://www.kcam.biz

ROBUST CROSS VALIDATIONS IN RIDGE REGRESSION

KANG-MO JUNG

ABSTRACT. The shrink parameter in ridge regression may be contaminated by outlying points. We propose robust cross validation scores in ridge regression instead of classical cross validation. We use robust location estimators such as median, least trimmed squares, absolute mean for robust cross validation scores. The robust scores have global robustness. Simulations are performed to show the effectiveness of the proposed estimators.

AMS Mathematics Subject Classification : 62J20, 62H99
Key words and phrases : Absolute mean, cross validation, least trimmed squares, median, ridge regression, shrink parameter.

1. Introduction

We consider the regression problem of predicting $y \in \mathbb{R}$ from independent variables $\mathbf{x} \in \mathbb{R}^p$, where the number of observations is n. When we take a linear form $\mu + \mathbf{x}^T\boldsymbol{\beta}$, where $\beta \in \mathbb{R}^p$ for predicting the response variable, it is called linear regression model. Ridge regression (Hoerl and Kennard, 1970) shrinks the parameter $\boldsymbol{\beta}$ by imposing a penalty on the size of coefficients. That is, we minimize over $\boldsymbol{\beta}$, a criterion of the form

$$\sum_{i=1}^{n}(y_i - \mu - \mathbf{x}_i^T\boldsymbol{\beta})^2 + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta} \tag{1}$$

for a regularized parameter $\lambda \in [0, \infty)$. Ridge regression is a regularized version of least squares regression using L_2 penalty on coefficient vector.

The criterion (1) uses a least squares method. It is well known that a least squares is very sensitive to a single outlier. Thus it needs a robust estimator of ridge regression (Jung, 2007). As λ ranges from 0 to ∞, the parameter $\boldsymbol{\beta}(\lambda)$ makes a path in $\mathbb{R}^p$. The parameter regulates the amount between penalty of regression parameter and residual sum of squares. The parameter λ should

Received October 2, 2008. Revised February 4, 2009. Accepted February 17, 2009.

be chosen by a disciplined way. It is obvious that we choose λ minimizes the mean squared error of β. There are so many results on the choice of the tuning parameter in nonparametric regression models as well as penalized regressions (Hastie et al., 2001). The cross validation method (Stone, 1974) is the most popular, which is to choose λ minimizing the classical cross validation (CCV) score function

$$CCV_\lambda = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_{(i)})^2, \tag{2}$$

where $\hat{y}_{(i)}$ is the estimate of the ith response variable based on the data omitting the ith observation. It is called 'leave-one-out' version of $\hat{y}_i$. Suppose that the ith observation was outlier. The estimate $\hat{y}_{(i)}$ is far away from y_i, so the score function (2) becomes some large. As least squares estimator is sensitive to outliers, so is the classical score function (2).

Wang and Scott (1994) proposed the absolute cross validation, and Park (2005) studied several robust score functions in nonparametric regressions. We can not find the study on robust score functions in ridge regression. In this article we propose robust score functions like the classical cross validation score function.

Section 2 gives several robust score functions in ridge regression. Classical cross validation used the sample mean. We proposed robust location estimators instead of the sample mean. And we describe the properties of the suggested robust score functions. Section 3 gives a small simulation to illustrate the effectiveness of the proposed estimator. Simulation results show that the proposed estimator seems to be more efficient than the classical cross validation when some of errors are contaminated. Also we can observe that the proposed method is robust for outlying observations. in ridge regression.

2. Robust score functions

The ridge regression solution for (1) can be written as the closed form

$$\hat{\boldsymbol{\beta}}^{ridge} = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y}. \tag{3}$$

This estimator is more or less stable than least squares estimator when the predictors has multicollinearity. The key problem is how to find the shrinkage parameter λ. There are many approaches for choosing the value of λ (see Meyers, 1986). The most popular criterion is cross validation in (2). Then by the result of Walker and Birch (1988) equation (2) becomes

$$CCV_\lambda = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - H_{ii}}\right)^2, \tag{4}$$

where H_{ii} is the ith diagonal element of the hat matrix $\mathbf{H}_\lambda = \mathbf{X}\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p\right)^{-1}\mathbf{X}^T$. Wahba et al. (1979) suggested the generalized cross validation

criterion (GCV) by substituting the denominator of (4) by $1 - \sum_i H_{ii}/n$. The GCV criterion can be used to solve a wide variety of problems.

Equation (4) is the sample mean for $z_i^2 = \Big((y_i - \hat{y}_i)/(1 - H_{ii})\Big)^2$. It is usual that the sample mean among location estimators is sensitive to outliers. We need a robust version of cross validation criterion to choose the tuning parameter in the ridge regression model. In particular even when the data has a single leverage point, we should use a robust version of cross validation. Because the leverage point has a large influence on the projection matrix $\mathbf{H}(\lambda)$.

We propose several robust versions of (4) as following

$$MEDCV_\lambda = \text{median } z_i^2, \tag{5}$$

$$TSSCV_\lambda = \frac{1}{h}\sum_{i=1}^{h} z_{i:n}^2, \tag{6}$$

$$MACV_\lambda = \frac{1}{n}\sum_{i=1}^{n} |z_i|, \tag{7}$$

where $z_{i:n}^2$ is the ith ascending order for z_i^2, $i = 1, ..., n$. Here h is the number of points for summation, and it is called coverage. When $h = n$ it becomes CCV. Equation (5) is the median cross validation criterion (MEDCV), (6) is the trimmed sum of squares cross validation criterion (TSSCV) and (7) is the mean absolute cross validation (MACV). They are also alternative robust estimators instead of the least squares estimator in linear regression (Rousseeuw and Leroy (1987)).

The proposed criterions are motivated by the fact that the terms in (2) are much influenced by outliers or influential observations. If the robust methods in (5) to (7) are used, the term $\hat{y}_{(i)}$ may be less influenced by observations having large residuals and leverages. Thus the cross validation criterions are little influenced by those observations.

However, the denominator of z_i can be influenced by observations having large residuals or leverages if the least squares estimator is used, because the least squares estimator is not robust. We should use a robust estimator instead of least squares estimator. Jung (2007) suggested a robust estimator for ridge regression using least trimmed squares estimator. And he also proposed an algorithm for getting the estimator. Instead of the objective function (1) we consider the sum of the smallest quantiles of squared residuals

$$\sum_{i=1}^{h} z_{(i:n)}^{*2} + \lambda\beta^T\beta, \tag{8}$$

where $z_i^{*2} = \Big(y_i - \mu - x_i^T\beta\Big)^2$ and the parameter λ is chosen by equations (5)-(7) with substituting z_i by z_i^*. Here the parenthesis subscript denotes the ascending order statistic.

The suggested methods for choosing the shrink parameter with the criterion (8) have several robustness properties.

Since the estimator satisfying (8) has high breakdown point for given λ, the suggested estimator has also high breakdown point. Formally the finite sample breakdown point (Donoho and Huber, 1983) is defined by

$$\epsilon_n^*(T) = \min\left\{\frac{m}{n}|sup_{X_m^n}||T(X_m^n) - T(X^n)|| = \infty\right\},$$

where $T(X^n)$ is an estimator for sample $X^n = \{X_1, ..., X_n\}$ and X_m^n is a sample by replacing m observations of X^n with arbitrary points. Roughly the breakdown point means the smallest fraction of a sample to make the estimator meaningless. For example the sample mean has $1/n$ breakdown point. Thus the sample mean is not robust.

The estimator from (8) is an $h-$sample estimate. When $h \approx n/2$, the estimator has a 50% breakdown point. Thus the suggested method for selecting λ has a 50% breakdown point. However the estimator by CCV has $1/n$ breakdown point. That is, it has asymptotically 0 % breakdown point.

3. Simulations

We performed a small simulation to show the robustness of the proposed estimator of the regularized parameter (Jung, 2007). We draw $m = 100$ samples of size $n = 20, 50, 100$ from the following model.

$$y_i = 1 + x_{1i} + x_{2i} + \cdots + x_{p-1,i} + \epsilon_i, p = 3, 6, 11,$$

where x_{ji}s for $j = 1, ..., p$ are distributed from $N(0, 100)$ and $\epsilon_i \sim N(0, 1)$. Here the correlation coefficients among predictors is 0.9. The ridge regression is designed for tackling the multicollinearity problem. This situation needs the ridge regression.

The simulation results are given in Table 1. The table shows that the suggested robust methods are little effective than CCV. Because the errors are normally distributed, so outliers have less influence on the regression coefficients. In this case the classical cross validation will be enough. However, the proposed methods are alternative methods to CCV when errors are coming from a normal distribution.

Next we performed suggested cross validations when errors are coming from t distribution with degrees of freedom 2. In this case the errors has a non-normal distribution. The simulation results are summarized in Table 2. It shows that the suggested robust methods are superior to CCV. Especially, MACCV is the most effective in most cases.

Finally we conducted four cross validations when 80% of cases are generated with $N(0, 1)$ errors, and 20% are contaminated by using an error $N(10, 1)$ We constructed outliers in the $y-$direction. The results are summarized in Table 3. This table shows that our proposed methods are superior to CCV. It implies

TABLE 1. Efficiencies of suggested methods and classical cross validation with normal errors.

$p\backslash n$		20	50	100
3	MEDCV	0.833	0.821	0.680
	TSSCV	0.774	0.639	0.683
	MACV	0.909	0.926	0.902
6	MEDCV	0.739	0.639	0.683
	TSSCV	0.794	0.774	0.758
	MACV	0.990	0.940	0.863
11	MEDCV	0.689	0.569	0.685
	TSSCV	0.881	0.730	0.873
	MACV	0.965	0.871	0.949

TABLE 2. Efficiencies of suggested methods and classical cross validation with t-distribution errors.

$p\backslash n$		20	50	100
3	MEDCV	0.863	0.758	0.830
	TSSCV	1.265	1.004	0.917
	MACV	1.458	1.030	0.989
6	MEDCV	0.632	0.708	0.885
	TSSCV	1.051	1.051	0.991
	MACV	1.445	1.106	1.001
11	MEDCV	1.494	0.870	0.904
	TSSCV	1.875	1.016	0.996
	MACV	1.863	1.033	0.998

that the suggested methods are robust from y-outliers. It shows that TSSCV is robust for the contaminated data, while CCV is sensitive to outliers.

In conclusion CCV provides good results if the errors are normally distributed, but then the proposed robust methods also behave well when the errors are normally distributed or contaminated.

4. Concluding remarks

Since the classical cross validation method in ridge regression is sensitive to outliers, it requires robust estimation. We used the least trimmed squares version of cross validation, median version and mean absolute version. The proposed estimator has a 50% breakdown point. Small simulations imply that the proposed estimates are robust and efficient.

TABLE 3. Efficiencies of suggested methods and classical cross validation with normal errors and 20% $y-$outliers.

$p\backslash n$		20	50	100
3	MEDCV	1.126	0.683	0.765
	TSSCV	2.052	1.222	1.113
	MACV	1.307	0.827	0.849
6	MEDCV	0.706	0.674	0.759
	TSSCV	1.053	1.186	1.014
	MACV	0.973	0.944	0.907
11	MEDCV	1.049	0.695	0.728
	TSSCV	1.581	1.107	1.013
	MACV	1.147	1.003	0.983

REFERENCES

1. D. L. Donoho and P. J. Huber, *The notion of breakdown point. In A Festschrift for Erich Lehmann, Edited by P. Bickel, K. Doksum and J. L. Hodges*, (1983), 157-184.
2. T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, 2001.
3. A. E. Hoerl and R. W. Kennard, *Ridge regression : biased estimation for nonorthogonal problems*, *Technometrics*, **12** (1970), 55-67.
4. K.-M. Jung, *A robust estimator in ridge regression*, *Journal of the Korean Data Analysis Society*, **9** (2007), 535-543.
5. R. H. Myers, *Classical and Modern Regression with Applications*, 2nd ed., Duxburry, Belmont, 1990.
6. D. Park, *Robust cross validation score*, *The Korean Communications in Statistics*, **12** (2005), 413-423.
7. P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*, John Wiley, New York, 1987.
8. M. Stone, *Cross-validatory choice and assessment of statistical predictions*, *J. of the Royal Statist. Soc., B*, **36** (1974), 111-147.
9. G. Wahba, G. H. Golub and C. G. Heath, *Generalized cross validation as a method for choosing a good ridge parameter*, *Technometrics*, **21** (1979), 215-223.
10. E. Walker and J. B. Birch, *Influence measures in ridge regression*, *Technometrics*, **30** (1988), 221-227.
11. F. Wang and D. Scott, *The L_1 method for robust nonparametric regression*, *J. of the Amer. Statist. Assoc.*, **89** (1994), 249-260.

Kang-Mo Jung received his BS from Seoul National University and Ph.D at KAIST. Since 1997 he has been a professor of Department of Informatics and Statistics at Kunsan National University. His research interests focus on robust statistics.

Department of Informatics and Statistics, Kunsan National University, Kunsan 573-701, Korea

e-mail: kmjung@kunsan.ac.kr