Journal of the Korean
Statistical Society
Vol. 22, No. 2, 1993

Optimal Fractions in terms of a Prediction-Oriented Measure

Wonwoo Lee [1]

ABSTRACT

The multicollinearity problem in a multiple linear regression model may present deleterious effects on predictions. Thus, it is desirable to consider the optimal fractions with respect to the unbiased estimate of the mean squares errors of the predicted values. Interestingly, the optimal fractions can be also illuminated by the Bayesian interpretation of the general James-Stein estimators.

KEYWORDS: Multicollinearity, MSE for Prediction, Fractional Principal Components Regression, Generalized James-Stein Estimator.

1. INTRODUCTION

The presence of multicollinearity in a multiple linear regression model has a number of potentially serious effects on the least squares(LS) estimates of the parameters. Thus, some alternative estimation techniques to the LS method have been introduced to remedy the problems caused by serious multicollinearity. Since the concepts of the alternatives are based upon shrinking the norm of the LS estimates of the coefficients, the resulting estimators become biased. Some of them are, for example, ridge estimator, principal components estimator, Stein-type estimator, Marquardt's fractional rank estimator, etc.

[1] Director, Lee's Statistical Analysis Lab, #401 Jeonwon Bldg., 16-47 Nonhyun-dong, Gangnam-gu, Seoul

However, it has been said that multicollinearity does not seem to necessarily have as deleterious effects on prediction as it does on the LS estimated coefficients because the quality of the prediction depends upon the location of the point at which one needs to predict. In other words, there are some regions in the regressor space where prediction will be effective and others where prediction will be quite poor due to the presence of severe multicollinearity. In fact, it is difficult to divide the good and bad regions for prediction.

Therefore, it is desirable to develop a biased estimator of the parameter vector from the viewpoint of prediction[Myers(1986,p.249)]. It can be done by minimizing the unbiased estimate of the mean squares errors(MSE) of the predicted values which is used as a performance criterion. Furthermore, it is shown that the new estimates of the coefficients are exactly equivalent to those provided by Bayesian interpretation of a general James-Stein estimator of the parameter vector.

The multicollinearity problem on prediction is reviewed in section 2. In section 3, the unbiased estimator of MSE of the predicted values is developed by utilizing the fractional principal components regression and the corresponding optimal fractions are derived. Section 4 is devoted to deriving Bayesian interpretation of a general James-Stein estimator. Finally, the concluding remarks and suggestions are in section 5.

2. EFFECTS OF MULTICOLLINEARITY ON PREDICTION

Consider the multiple linear regression(MLR) model,

$$\underline{y} = X \underline{\beta} + \underline{\epsilon}, \tag{2.1}$$

where $\underline{y}$ is an $(n \times 1)$ vector of responses, X is an $(n \times p)$ matrix of the predetermined, standardized regressors, $\underline{\beta}$ is a $(p \times 1)$ parameter vector, and $\underline{\epsilon}$ is an $(n \times 1)$ random error vector with $E(\underline{\epsilon}) = \underline{0}$ and $Var(\underline{\epsilon}) = \sigma^2 I$. Faced with the multicollinearity problem in (2.1), a reparameterized model based on the eigenvalue decomposition will be useful for the analysis:

$$\underline{y} = Z \underline{\alpha} + \underline{\epsilon}, \tag{2.2}$$

where $Z = XV$, $\underline{\alpha} = V' \underline{\beta}$, and V is the $(p \times p)$ matrix of the eigenvectors of $X'X$, i.e., $V = [\underline{v}_1, \ldots, \underline{v}_p]$, $\underline{v}_j$ is the jth eigenvector of $X'X$. Note that V is an orthogonal matrix and $Z'Z = \Lambda = diag(\lambda_1, \ldots, \lambda_p)$, where the λ_j is the jth largest eigenvalue of $X'X$.

If one wants to predict at the point, $\underline{x}_o$, then the quality of prediction can be measured by the variance of $\hat{y}(\underline{x}_o)$, apart from σ^2, i.e.,

$$\begin{aligned}\sigma^{-2}Var(\hat{y}(\underline{x}_o)) &= \underline{x}_o'(X'X)^{-1}\underline{x}_o \\ &= \underline{x}_o'V\Lambda^{-1}V'\underline{x}_o \\ &= \underline{z}_o'\Lambda^{-1}\underline{z}_o \\ &= \sum_{j=1}^{p} z_{oj}^2/\lambda_j. \end{aligned} \tag{2.3}$$

Recalling that if multicollinearity is present in the model, then at least one $\lambda_j \cong 0$, when the corresponding $z_{oj} = \underline{x}_o'\underline{v}_j$ to $\lambda_j \cong 0$ is very close to zero, the prediction at $\underline{x}_o$ will work well. Otherwise, the prediction will be quite poor. Thus, the location of the point to predict determines the quality of prediction.

Therefore, as far as the prediction is concerned with the multicollinearity problem, the mean squares errors of the predicted values(MSEP) can be an interesting measure for the performance of the multiple linear regression model. Since MSEP is unknown, its unbiased estimate will be utilized in this paper. There are similar prediction criteria for ridge regression[Wahba et.al.(1979), Myers(1986)].

3. OPTIMAL FRACTIONS WITH RESPECT TO MSEP

Consider the fractional principal components(FPC) regression as follows[Lee(1986)];

$$\begin{aligned}\underline{y} &= Z\underline{\alpha} + \underline{\epsilon} \\ &= ZF^{-}F\,\underline{\alpha} + \underline{\epsilon} \\ &= Z_F\,\underline{\alpha}_F + \underline{\epsilon},\end{aligned}$$

where $F = diag(f_1, \ldots, f_p)$, $0 \le f_i \le 1$ for all i, $Z_F = ZF^{-}$, and $\underline{\alpha}_F = F\underline{\alpha}$. The diagonal matrix F is termed the fraction matrix and the diagonals f_j are named as the fractions. Note that F^{-} is a generalized inverse of F. Then the LS estimator of $\underline{\alpha}_F$ can be obtained as of the form,

$$\begin{aligned}\hat{\underline{\alpha}}_{F,LS} &= (Z_F'Z_F)^{-1}Z_F'\underline{y} \\ &= F\hat{\underline{\alpha}}_{LS}.\end{aligned} \tag{3.1}$$

The expression in (3.1) is the general form of the biased estimators of the coefficients, α_j's, for combatting multicollinearity. In other words, the fractions, f_j, $j = 1, \ldots, p$, may take different set of values for various estimation techniques. For example, (1)(Principal Components Regression) $f_1 = \ldots = f_r = 1$, $f_{r+1} = \ldots = f_p = 0$ (2)(Ridge Regression) $f_j = \lambda_j/(\lambda_j + k)$, $j = 1, \ldots, p$ (3)(Stein) $f_j = 1 - c\sigma^2/(\hat{\underline{\alpha}}_{LS}'\Lambda\hat{\underline{\alpha}}_{LS})$ (4)(Fractional Rank) $f_1 = \ldots = f_r = 1$, $f_{r+1} = \lambda_{r+1}\alpha_{r+1}^2/(\lambda_{r+1}\alpha_{r+1}^2 + \sigma^2)$, $f_{r+2} = \ldots = f_p = 0$,etc.

Using the biased estimator $\underline{\hat{\alpha}}_F$ because of multicollinearity, the fitted values are

$$\begin{aligned}\hat{\underline{y}}_F &= Z\underline{\hat{\alpha}}_F = ZF(Z'Z)^{-1}Z'\underline{y} \\ &= H_F\underline{y},\end{aligned} \tag{3.2}$$

where $H_F = ZF(Z'Z)^{-1}Z'$. Then the MSE of $\hat{\underline{y}}_F$ can be expressed as follows;

$$\begin{aligned}MSE(\hat{\underline{y}}_F) &= \sum_{i=1}^{n} MSE(\hat{y}_{i,F}) \\ &= \sum_{i=1}^{n} Var(\hat{y}_{i,F}) + \sum_{i=1}^{n} Bias^2(\hat{y}_{i,F})\end{aligned}$$

In order to obtain the unbiased estimate of $MSE(\hat{\underline{y}}_F)$, named as UMSEP, first, look at

$$\begin{aligned}Var(\hat{y}_{i,F}) &= Var(\underline{x}_i'\underline{\hat{\beta}}_F) = Var(\underline{x}_i'V\underline{\hat{\alpha}}_F) \\ &= \sigma^2\underline{x}_i'VF(Z'Z)^{-1}Z'Z(Z'Z)^{-1}FV'\underline{x}_i \\ &= \sigma^2\underline{x}_i'VF^2\Lambda^{-1}V'\underline{x}_i.\end{aligned}$$

Thus,

$$\begin{aligned}\sum_{i=1}^{n} Var(\hat{y}_{i,F}) &= \sigma^2 tr(XVF^2\Lambda^{-1}V'X') \\ &= \sigma^2 tr(ZF^2\Lambda^{-1}Z') \\ &= \sigma^2 tr(H_F^2) = \sigma^2 tr(F^2)\end{aligned} \tag{3.3}$$

Therefore, the unbiased estimate of the 'variance part' is $s^2 tr(H_F^2)$, where s^2 is the residual mean squares.

Secondly, the 'squared-bias part' is reexpressed as

$$\begin{aligned}\sum_{i=1}^{n} Bias^2(\hat{y}_{i,F}) &= (X\underline{\beta}-E(X\underline{\hat{\beta}}_F))'(X\underline{\beta}-E(X\underline{\hat{\beta}}_F)) \\ &= (X\underline{\beta}-XVF\Lambda^{-1}Z'Z\underline{\alpha})'(X\underline{\beta}-XVF\Lambda^{-1}Z'Z\underline{\alpha}) \\ &= (X\underline{\beta})'(I-H_F)^2(X\underline{\beta}).\end{aligned} \tag{3.4}$$

The unbiased estimate of the term in (3.4) can be established by considering

$$\begin{aligned}E(SSE_F) &= E((\underline{y}-X\underline{\hat{\beta}}_F)'(\underline{y}-X\underline{\hat{\beta}}_F)) \\ &= E((\underline{y}-ZF\Lambda^{-1}Z'\underline{y})'(\underline{y}-ZF\Lambda^{-1}Z'\underline{y})) \\ &= E(\underline{y}'(I-H_F)^2\underline{y}) \\ &= (X\underline{\beta})'(I-H_F)^2(X\underline{\beta}) + \sigma^2 tr[(I-H_F)^2].\end{aligned}$$

So, the unbiased estimate of the 'squared-bias part' is $SSE_F - s^2 tr[(I-H_F)^2]$.

Consequently, the unbiased estimate of $MSE(\hat{\underline{y}}_F)$ is

$$\begin{aligned} UMSEP &= s^2 tr(H_F^2) + SSE_F - s^2 tr[(I - H_F)^2] \\ &= SSE_F + s^2(2tr(H_F) - n). \end{aligned} \tag{3.5}$$

Note that, for the special case, $F = I$, it turns out to be ps^2 which is the unbiased estimate of $\sum Var(\hat{y}_i)$.

The fractions, which are the diagonal elements of F, play important roles in this FPC framework because they determine the shrinking proportions of the LS estimated parameters, $\hat{\alpha}_j$, $j = 1, \ldots, p$, which are inflated due to multicollinearity. The optimal values of the f_j's, in terms of a prediction criterion, are nothing but those for which $UMSEP$ is minimized.

The first derivatives of $UMSEP$ with respect to the f_j's can be expressed in the matrix form with defining

$$\frac{\partial UMSEP}{\partial F} = diag\left(\frac{\partial UMSEP}{\partial f_1}, \ldots, \frac{\partial UMSEP}{\partial f_p}\right)$$

i.e.,

$$\begin{aligned} \frac{\partial UMSEP}{\partial F} &= 2s^2 \frac{\partial tr(H_F)}{\partial F} + \frac{\partial SSE_F}{\partial F} \\ &= 2s^2 I + (-2\frac{\partial \underline{y}' H_F \underline{y}}{\partial F} + \frac{\partial \underline{y}' H_F^2 \underline{y}}{F}) \\ &= 2s^2 I - 2 \begin{bmatrix} \lambda_1 \hat{\alpha}_1^2 & & \\ & \ddots & \\ & & \lambda_p \hat{\alpha}_p^2 \end{bmatrix} + 2 \begin{bmatrix} f_1 \lambda_1 \hat{\alpha}_1^2 & & \\ & \ddots & \\ & & f_p \lambda_p \hat{a}lpha_p^2 \end{bmatrix} \end{aligned} \tag{3.6}$$

Therefore, the optimal values of the f_j's in terms of UMSEP are

$$f_j = 1 - \frac{s^2}{\lambda_j \hat{\alpha}_j^2}, \quad j = 1, \ldots, p. \tag{3.7}$$

Note that the Hessian matrix is positive definite.

Interestingly, these optimal values can be illuminated from the viewpoint of the Bayesian interpretation of a generalized James-Stein estimator which will be derived in the next section.

4. BAYESIAN-INTERPRETED GENERALIZED JAMES-STEIN ESTIMATOR

The concept of shrinking the LS estimator, proposed by Stein, has been justified as a broad class of alternative precedures to LS estimation when multicollinearity is present in the MLR model. The Stein estimator for $\underline{\hat{\alpha}}, \underline{\hat{\alpha}}_{ST}$, is

$$\underline{\hat{\alpha}}_{ST} = (1 - \frac{c\sigma^2}{\underline{\hat{\alpha}}'_{LS}\Lambda\underline{\hat{\alpha}}_{LS}})\underline{\hat{\alpha}}_{LS}. \tag{4.1}$$

Note that $MSE(\underline{\hat{\alpha}}_{ST}) < MSE(\underline{\hat{\alpha}}_{LS})$ for $0 < c < 2(p-2)$. Furthermore, substituting s^2 for σ^2, the resulting estimator, so-called the James-Stein(JS) estimator, is

$$\underline{\hat{\alpha}}_{JS} = (1 - \frac{cs^2}{\underline{\hat{\alpha}}'_{LS}\Lambda\underline{\hat{\alpha}}_{LS}})\underline{\hat{\alpha}}_{LS}. \tag{4.2}$$

By the way, it is worth while to look at its Bayesian interpretation and the derivation of the JS estimator [Vinod and Ullah(1981)]. Assuming that

$$\begin{aligned} &(A1)\ \underline{y} \sim N(Z\underline{\alpha}, \sigma^2 I) \\ &(A2)\ \underline{\alpha} \sim N(\underline{\alpha}_o, \sigma^2_\alpha (Z'Z)^{-1}) \end{aligned} \tag{4.3}$$

under the quadratic loss function, the Bayes estimator of $\underline{\alpha}$ which is the posterior mean vector, is

$$\underline{\hat{\alpha}}_B = (\sigma^{-2}Z'Z + \sigma_\alpha^{-2}Z'Z)^{-1}(\sigma^{-2}Z'Z\underline{\hat{\alpha}}_{LS} + \sigma_\alpha^{-2}Z'Z\underline{\alpha}_o).$$

Letting, now, $\underline{\alpha}_o = \underline{0}$ (it implies to shrink toward to zero),

$$\begin{aligned} \underline{\hat{\alpha}}_B &= \frac{\sigma^2_\alpha}{\sigma^2 + \sigma^2_\alpha}\underline{\hat{\alpha}}_{LS} \\ &= (1 - \frac{\sigma^2}{\sigma^2 + \sigma^2_\alpha})\underline{\hat{\alpha}}_{LS}. \end{aligned} \tag{4.4}$$

Replacing the term $(\sigma^2 + \sigma^2_\alpha)$ by its unbiased estimator based upon the LS estimation [2], $\underline{\hat{\alpha}}'_{LS}\Lambda\underline{\hat{\alpha}}_{LS}/p$, the Bayesian-interpreted Stein estimator of $\underline{\alpha}$ can be expressed as

$$\underline{\hat{\alpha}}_{BS} = (1 - \frac{p\sigma^2}{\underline{\hat{\alpha}}'_{LS}\Lambda\underline{\hat{\alpha}}_{LS}})\underline{\hat{\alpha}}_{LS}, \tag{4.5}$$

[2]
$$\begin{aligned} E(\underline{\hat{\alpha}}'_{LS}\Lambda\underline{\hat{\alpha}}_{LS}) &= E[(\underline{\alpha} + (Z'Z)^{-1}Z'\underline{\epsilon})'\Lambda(\underline{\alpha} + (Z'Z)^{-1}Z'\underline{\epsilon})] \\ &= E(\underline{\alpha}'\Lambda\underline{\alpha}) + E(\underline{\epsilon}'Z\Lambda^{-1}Z'\underline{\epsilon}) + 2E(\underline{\alpha}'Z'\underline{\epsilon}) \\ &= p\sigma^2_\alpha + p\sigma^2. \end{aligned}$$

which is equivalent to the Stein estimator in (4.1) by taking $c = p$. In addition, substituting σ^2 by s^2, the Bayesian-interpreted James-Stein(JS) estimator of $\underline{\alpha}$ is

$$\hat{\underline{\alpha}}_{BJS} = (1 - \frac{ps^2}{\hat{\underline{\alpha}}'_{LS}\Lambda\hat{\underline{\alpha}}_{LS}})\hat{\underline{\alpha}}_{LS}. \tag{4.6}$$

The Bayesian-interpreted JS estimator seems to be practical, but it utilizes a uniform shrinkage which is improper in the context of the multicollinearity preblem. Therefore, by assuming a more general prior density of $\underline{\alpha}$, a more sensible Bayesian interpretation can be made [Lee(1986)].

Instead of (A2) in (4.3), the prior density of $\underline{\alpha}$ can be postulated as follows;

(GA2) $\underline{\alpha} \sim N(\underline{\alpha}_o, D_\alpha(Z'Z)^{-1})$,

where $D_\alpha = diag(\sigma^2_{\alpha_1}, \ldots, \sigma^2_{\alpha_p})$.

The prior density in (GA2) looks more general than that in (A2), because it is more meaningful to assume that the variances of the α_j's depend upon not only the eigenvalues but also the self-endowed values, the σ_{α_j}'s. With the assumptions (A1) and (GA2), the Bayes estimator of $\underline{\alpha}$, labelled by GB, is

$$\begin{aligned}\hat{\underline{\alpha}}_{GB} &= (\sigma^{-2}Z'Z + D_\alpha^{-1}Z'Z)^{-1}(\sigma^{-2}Z'Z\hat{\underline{\alpha}}_{LS} + D_\alpha^{-1}Z'Z\underline{\alpha}_o) \\ &= (\sigma^{-2}I + D_\alpha^{-1})^{-1}(\sigma^{-2}\hat{\underline{\alpha}}_{LS} + D_\alpha^{-1}\underline{\alpha}_o).\end{aligned}$$

That is,

$$\hat{\alpha}_{j,GB} = (\frac{\sigma^2_{\alpha_j}}{\sigma^2 + \sigma^2_{\alpha_j}})\hat{\alpha}_j + (\frac{\sigma^2}{\sigma^2 + \sigma^2_{\alpha_j}})\alpha_{j,0}, \text{ for all } j.$$

Now, taking $\alpha_{j,0} = 0,\ j = 1, \ldots, p$,

$$\hat{\alpha}_{j,GB} = (1 - \frac{\sigma^2}{\sigma^2 + \sigma^2_{\alpha_j}})\hat{\alpha}_j. \tag{4.7}$$

The results in (4.7) are parallel to those in (4.4). In order to obtain the unbiased estimator of $(\sigma^2 + \sigma^2_{\alpha_j})$, defining $\Lambda_j(0) = diag(0, \ldots, 0, \lambda_j, 0, \ldots, 0)$,

$$\begin{aligned}E(\lambda_j\hat{\alpha}_j^2) &= E(\hat{\underline{\alpha}}'_{LS}\Lambda_j(0)\hat{\underline{\alpha}}_{LS}) \\ &= E((\Lambda^{-1}Z'(Z\underline{\alpha}+\underline{\epsilon}))'\Lambda_j(0)(\Lambda^{-1}Z'(Z\underline{\alpha}+\underline{\epsilon})) \\ &= E(\underline{\alpha}'\Lambda_j(0)\underline{\alpha}) + E(\underline{\epsilon}'Z\Lambda^{-1}\Lambda_j(0)\Lambda^{-1}Z'\underline{\epsilon}) \\ &= tr(D_\alpha\Lambda^{-1}\Lambda_j(0)) + tr(\sigma^2\Lambda^{-1}\Lambda_j(0)) \\ &= \sigma^2_{\alpha_j} + \sigma^2\end{aligned} \tag{4.8}$$

Thus, from(4.8), the unbiased estimator of $(\sigma^2_{\alpha_j} + \sigma^2)$ is $\lambda_j\hat{\alpha}_j^2$.

Plugging this result into (4.7) and, in addition, substituting s^2 for σ^2, the generalized Bayesian-interpreted James-Stein (GBJS) estimators of the α_j's are developed as

$$\hat{\alpha}_{j,GBJS} = (1 - \frac{s^2}{\lambda_j \hat{\alpha}_j^2})\hat{\alpha}_j, \quad j = 1, \ldots, p. \tag{4.9}$$

It is interesting to investigate that the estimated shrinkage values for the GBJS estimators are equivalent to the optimal fractions based on UMSEP in Sec. 3. Note that the quadratic loss function in terms of $\underline{\alpha}$ is selected for obtaining the GBJS estimator [Baranchik(1970)].

5. CONCLUDING REMARKS AND SUGGESTIONS

When multicollinearity is dipped into the MLR model, one of the alternative methods to the LS method, the Stein-type estimator can be chosen to shrink the LS estimator. However, the optimal fractions with respect to the unbiased estimate of the mean squares errors of the predicted values have the same forms as the generalized Stein-type solutions which utilize the different shrinkages. In other words, the generalized JS estimators can be recommended if the aim is to get estimators with the 'better' MSE of the predicted values.

Some extensive simulation studies can be suggested for comparisons among the prediction-oriented measures.

REFERENCES

(1) Baranchik, A.J.(1970). A Family of Minimax Estimators of the Mean of a Multivariate Normal Distribution. *Annals of Mathematical Statistics*, Vol.41, 642-645.

(2) James, W. and Stein,C.(1961). Estimation with Quadratic Loss. *Proceedings of Fourth Berkeley Symposium in Mathematical Statistics and Probability*, Vol.1, 361-379.

(3) Lee, W.(1986). *Fractional Principal Components Regression: A General Approach to Biased Estimators.* Unpublished Ph.D. dissertation, Dept. of Statistics, VPI & SU.

(4) Marquardt, D.W.(1970). Generalized Inverse, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation. *Technometrics*, Vol.12, 591-612.

(5) Myers, R.H.(1986). *Classical and Modern Regression with Applications.* Boston: Duxbury Press.

(6) Vinod, H.D. and Ullah, A.(1981). *Recent Advances in Regression Method.* New York and Basel: Marcel Dekker, Inc.

(7) Wahba, G., Gloub, G.H. and Heath, C.G.(1979). Generalized Cross Validation as a Method for Choosing a Good Ridge Parameter. *Technometrics*, Vol.21, 215-223.