OPTIMAL PREDICTION IN LINEAR REGRESSION ANALYSIS

by

Janella F. Pantula and Lawrence L. Kupper

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1872

January 1990

Optimal Prediction in Linear Regression Analysis

Janella F. Pantula
Research Triangle Institute
Research Triangle Park, NC 27709

Lawrence L. Kupper
Department of Biostatistics
University of North Carolina
Chapel Hill, NC 27514

Expressions are derived for generalized ridge and ordinary ridge predictors that are optimal in terms of mean squared error of prediction (MSEP) for predicting the response at a single or at multiple future observation(s). Using the MSEP criterion, operational predictors are compared to the ordinary least squares (OLS) predictor and to several biased predictors derived from some popular biased estimators. Simulation results indicate that the performance of these predictors depends on the direction of the prediction, the magnitude of the signal-to-noise ratio, the level of multicollinearity, and the number of explanatory variables.

KEY WORDS: mean squared error of prediction, biased estimation, ridge regression

1. INTRODUCTION

Consider the linear regression model

$$y_{nx1} = X_{nxr}\,\beta_{rx1} + \epsilon_{nx1},$$

where β is an rx1 vector of unknown regression coefficients, y is an nx1 vector of sample mean corrected observable responses, X is a full rank model matrix of fixed explanatory variables that are centered and scaled, ϵ is an nx1 vector of random errors such that $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2 I$.

Note that the form of the X matrix is such that $X'X$ is a correlation matrix.

It is always possible to write the usual linear regression model in canonical form. In particular, there exists an rxr orthogonal matrix P such that $X'X = P\Lambda P'$, where Λ is the diagonal matrix of ordered eigenvalues of $X'X$, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r > 0$, P is such that $PP' = P'P = I_{rxr}$, and the columns of P are the corresponding normalized eigenvectors of $X'X$. Let $X = ZP'$; then we have

$$y_{nx1} = Z_{nxr}\, \alpha_{rx1} + \epsilon_{nx1},$$

where $\alpha = P'\beta$, $Z'Z = \Lambda$, and $\alpha'\alpha = \beta'PP'\beta = \beta'\beta$.

Further, consider the linear regression model used in the prediction of future observations, namely,

$$\begin{aligned} y^*_{mx1} &= X^*_{mxr}\, \beta_{rx1} + \epsilon^*_{mx1} \\ &= Z^*_{mxr}\, \alpha_{rx1} + \epsilon^*_{mx1}. \end{aligned}$$

Here, y^* is an mx1 vector of future response observations, $Z^* = X^*P$ is an mxr matrix of transformed fixed future explanatory variables, ϵ^* is an mx1 vector of random errors corresponding to the prediction of y^*. We assume that $E(\epsilon^*) = 0$, $Var(\epsilon^*) = \sigma^2 I_{mxm}$, and $Cov(\epsilon, \epsilon^*) = 0$; i.e., the random errors from the usual regression model are uncorrelated with the random errors from the prediction regression model.

Define the predictor of y^* as

$$\begin{aligned} \tilde{y}^*_{mx1} &= X^*_{mxr}\, \tilde{\beta}_{rx1} \\ &= Z^*_{mxr}\, \tilde{\alpha}_{rx1} \end{aligned}$$

where $\tilde{\alpha} = P'\tilde{\beta}$ is some estimator of the unknown vector of transformed regression coefficients α.

Here, we are concerned with the prediction of y^* as opposed to the estimation of α. We define the term Mean Squared Error of Prediction (MSEP) only for forecasting the unobservable future response vector y^*. Under the given assumptions, we have

$$\begin{aligned} MSEP(\tilde{y}^*) &= m\sigma^2 + tr\,[Var(\tilde{y}^*)] + [bias(\tilde{y}^*)]'[bias(\tilde{y}^*)] \\ &= m\sigma^2 + MSE(\tilde{y}^*)\ , \end{aligned}$$

where $MSE(\tilde{y}^*) = E[(\tilde{y}^* - Z^*\alpha)'(\tilde{y}^* - Z^*\alpha)]$.

Let $\hat{\alpha} = A^{-1}X'y$; then, $\hat{y}^* = Z^*\hat{\alpha}$ is the Ordinary Least Squares (OLS) predictor. Thus,

$$\begin{aligned} MSEP(\hat{y}^*) &= m\sigma^2 + tr\,[Var(\hat{y}^*)] \\ &= \sigma^2\,(m + \Sigma_{j=1}^{r}\ 1/\lambda_j\ \Sigma_{i=1}^{m}\ z_{ij}^{*2}). \end{aligned}$$

It is well known in the estimation literature that the behavior of the OLS estimator is sometimes unacceptable due to the variance inflation possibly caused by near multicollinearity in the explanatory variables. Upon examination of the $MSEP(\hat{y}^*)$, one can see that the behavior of the OLS predictor is affected by one or more small eigenvalues of $X'X$, as well by the elements of the matrix Z^* of transformed explanatory variables.

Several alternatives to the OLS estimator have been recommended in an estimation context, which, under the assumed model, are biased. However, these alternatives may be preferable to the least squares estimator if some criteria other than simple unbiasedness (eg. Mean Squared Error) is used. We have derived alternative predictors which are optimal if the MSEP criterion is adopted.

2. OPTIMAL PREDICTORS

In this discussion, we focus on biased estimation in a prediction context. In particular, a Generalized Ridge (GR) estimator, and the corresponding predictor, are derived based on a minimum MSEP criterion. Also, an Ordinary Ridge (OR) predictor, based on an estimator similar to that of Hoerl, Kennard and Baldwin's (1975) estimator, is developed. The GR and OR predictors will involve $\tilde{\alpha}$ which depends both on the original data and on Z^*.

Let

$$\tilde{y}_K^* = Z^* K \hat{\alpha} \quad , \tag{2.1}$$

where K is a diagonal matrix with elements k_i, i=1,...,r. For $\tilde{y}_K^*$ in (2.1), we have

$$MSEP(\tilde{y}_K^*) = \sigma^2[m + tr\,(Z^{*'}Z^*K^2\,\Lambda^{-1})] + \alpha'(I-K)'Z^{*'}Z^*(I-K)\alpha \; . \tag{2.2}$$

The Generalized Ridge (GR) predictor, based on the estimator of Hoerl and Kennard (1970), is given by (2.1) with $K = K_{GR} = (\Lambda + D)^{-1}\,\Lambda$, where D is an rxr diagonal matrix. Similarly, for the Ordinary Ridge (OR) predictor, based on the OR estimator of Hoerl and Kennard (1970), the matrix $K = K_{OR} = (\Lambda + c_{OR}I)^{-1}\,\Lambda$, where $c_{OR} \geq 0$ is a constant.

We now present three important results, the proofs of which can be found in the Appendix.

Result 1: The proposed GR predictor based on minimizing the MSEP of the predictor with respect to the diagonal elements of D (or K) yields

$$\tilde{K} = diag\,\{(A^{-1}b)'\}, \tag{2.3}$$

where

$$Z^{*'}Z^* = H = ((h_{ij})),$$

$$A = ((a_{ij})) ,$$

$$a_{ij} = \begin{cases} h_{ij}\,\alpha_i\,\alpha_j & i \neq j \\ h_{ij}\,(\alpha_i^2 + \sigma^2/\lambda_i) & i = j, \end{cases}$$

and

$$b_i = \alpha_i \{h_{i1}\, h_{i2} \cdots h_{ir}\}\, \alpha ,$$

for i=1,2,...,r and j=1,2,...,r. Therefore, $\tilde{K}_{GR} = \tilde{K} = (A + \tilde{D})^{-1}A$, where $\tilde{d}_i = \lambda_i(1 - \tilde{k}_i)/\tilde{k}_i$, and $\tilde{k}_i$ is the ith diagonal element of $\tilde{K}$ for i=1,2,...,r.

Similar results were obtained by Obenchain (1978).

Result 2: If m = 1 (i.e., if we are interested in predicting only one observation), then (2.3) simplifies to

$$\tilde{k}_i = \alpha_i\lambda_i\, (\Sigma_{j=1}^{r} z_j^* \alpha_j) \;/\; [z_i^*(\sigma^2 + \Sigma_{j=1}^{r} \alpha_j^2 \lambda_j)] , \qquad (2.4)$$

for i = 1,2,...,r.

Result 3: Suppose X'X = I. Then, the c_{OR} that minimizes $MSEP(\tilde{y}_{OR}^*)$ is

$$\tilde{c}_{OR} = \sigma^2 tr\,[Z^{*\prime}Z^*] \;/(\alpha' Z^{*\prime} Z^* \alpha) . \qquad (2.5)$$

Hoerl, Kennard, and Baldwin (1975) proposed an operational OR estimator by choosing c_{OR} such that $MSE(\tilde{\beta}_{OR})$ is minimized for X'X = I.

Consider predicting one future observation $y^* = z^{*\prime}\alpha + \epsilon^*$; then $MSEP(\tilde{y}_{OR}^*)$ is minimized with respect to c_{OR} when

$$\tilde{c}_{OR} = \sigma^2(\Sigma_{j=1}^{r} z_j^{*2}) \;/\; (\Sigma_{j=1}^{r} z_j^* \alpha_j)^2 . \qquad (2.6)$$

Our optimal predictors that minimize $MSEP(\tilde{y}^*)$ clearly depend on the population values of the regression coefficients and on the value of σ^2. Thus, we suggest the use of operational GR and OR predictors obtained by substituting OLS estimates of α and σ^2 in expressions (2.3) and (2.5).

3. OPERATIONAL PREDICTORS

The operational GR predictor, $\hat{y}^*_{GPK}$ based on the minimum MSEP($\tilde{y}^*_{GR}$) predictor is found by substituting OLS estimators, $\hat{\alpha}$ and s^2, for α and σ^2 in equation (2.3).

The single prediction operational GR predictor corresponding to equation (2.4) simplifies to

$$\hat{k}_{GPKi} = \hat{\alpha}_i\lambda_i(\Sigma_{j=1}^r z_j^*\hat{\alpha}_j) \; / \; [z_i^*(s^2 + \Sigma_{j=1}^r \hat{\alpha}_j^2\lambda_j)] \; ,$$

for $i = 1,2,\ldots,r$.

Result 4: In the case of predicting one future observation, the predictor $\hat{y}^*_{GPK} = z^{*\prime}\hat{K}_{GPK}\hat{\alpha}$ is a shrunken predictor of $\hat{y}^* = z^{*\prime}\hat{\alpha}$, the OLS predictor. (See the Appendix for a proof.)

The operational OR predictor, $\hat{y}^*_{OPK}$, based on the minimum MSEP($\tilde{y}^*_{OR}$) predictor is found by substituting OLS estimators, $\hat{\alpha}$ and s^2, for α and σ^2 in equation (2.5), obtaining

$$\hat{c}_{OPK} = s^2 \; tr \; [Z^{*\prime}Z^*] \; / \; \hat{y}^{*\prime}\hat{y}^* .$$

This simplifies to

$$\hat{c}_{OPK} = s^2 \; (\Sigma_{j=1}^r z_j^{*2}) \; / \; \hat{y}^{*2},$$

when predicting one future observation.

Explicit expressions for the bias and the mean squared error of prediction for the operational predictors proposed in this section and in the existing literature are difficult to obtain. Thus, the properties of the OLS predictor, of the existing GR and OR predictors, and of the newly proposed operational predictors will be explored using Monte Carlo methods.

4. SIMULATION STUDY

The major difference between previous simulation studies and the one under consideration here is our emphasis on prediction at particular future observations.

We have investigated the relative performance of six predictors: two predictors based on the MSEP optimality criterion, three predictors based on the biased estimation literature, and the OLS predictor. We calculated the predictors for each single prediction and for multiple predictions. The six predictors are as follows:

1. OLS - The Ordinary Least Squares predictor, $\hat{y}^* = Z^*\hat{\alpha}$.

2. GPK - The operational GR predictor based on the minimum MSEP criterion introduced in Section 3.

3. GHK - The operational GR predictor based on the minimum MSE criterion as introduced by Hoerl and Kennard (1970), which can be written as $\hat{y}^*_{GHK} = Z^*\hat{K}_{GHK}\hat{\alpha}$, where $\hat{K}_{GHK}$ is a diagonal matrix with elements $\hat{\alpha}_i^2\lambda_i / (\hat{\alpha}_i^2\lambda_i + s^2)$, $i=1,\ldots,r$.

4. OPK - The operational OR predictor based on the minimum MSEP criterion introduced in Section 3.

5. OHK - The iterated version of the minimum MSE operational OR predictor due to Hoerl and Kennard (1976), which can be written as $\hat{y}^*_{OHK} = Z^*\hat{K}_{OHK}\hat{\alpha}$, where $\hat{K}_{OHK}$ is a diagonal matrix with elements $\lambda_i / (\lambda_i + \hat{c}_{OHK})$, $i=1,2,\ldots,r$; $\hat{c}_{OHK}$ is determined by an iterative algorithm with initial $\hat{c}_{OHK,0} = rs^2 / \hat{\alpha}'\hat{\alpha}$. Our stopping rule is as follows: let $\delta = 20T^{-1.30} \geq 0$, where $T = tr(X'X)^{-1}/r$ (Hoerl and Kennard, 1976). If $\hat{\alpha}(\hat{c}_{OHK,i})'\hat{\alpha}(\hat{c}_{OHK,i}) < (1+\delta)^{-1}\,\hat{\alpha}(\hat{c}_{OHK,i-1})'\hat{\alpha}(\hat{c}_{OHK,i-1})$, where $\hat{c}_{OHK,i}$ is the

is the value of $\hat{c}_{OHK}$ in the i-th iteration, then continue; or else terminate and use $\hat{\alpha}(\hat{c}_{OHK,i})$. If $\hat{c}_{OHK}$ does not converge in five iterations, then we take $\hat{c}_{OHK}$ to be $\hat{c}_{OHK,0}$.

6. OLW - The operational OR predictor proposed by Lawless and Wang (1976), which can be written as $\hat{y}^*_{OLW} = Z^*\hat{K}_{OLW}\hat{\alpha}$, where $\hat{K}_{OLW}$ is a diagonal matrix with elements $\lambda_i/(\lambda_i + \hat{c}_{OLW})$, i=1,2,...,r, and where $\hat{c}_{OLW} = rs^2/(\hat{\alpha}'\Lambda\hat{\alpha})$.

Mainly, two comparison criteria have been considered in the literature: Mean Squared Error (MSE($\tilde{\beta}$)) and Predictive Mean Squared Error (PMSE($\tilde{\beta}$)). Note that PMSE($\tilde{\beta}$) = MSE($\tilde{y}$), where $\tilde{y} = X\tilde{\beta}$. Lawless and Wang (1976), Vinod (1976), Dempster, Schatzoff, and Wermuth (1977), and Hoerl, Schuenemeyer, and Hoerl (1986) have considered both the PMSE and MSE criteria. Others have considered only the MSE criterion. We use the MSEP criterion to compare our operational predictors with the OLS predictor and with other operational predictors based on biased estimators. The MSEP criterion is the most appropriate criterion for comparison, since, in most cases, one is estimating the regression coefficients with the intended purpose of prediction. This criterion has not previously been considered for comparing the traditional ordinary ridge predictors in a simulation study. Copas (1983) used the MSEP criterion to compare a shrunken least square predictor to the OLS predictor.

Dempster, Schatzoff, and Wermuth (1977) commented that the connection between the PMSE criterion and prediction is as follows: if $\hat{y} = X\hat{\beta}$ is used to predict $y^* = X\beta + \epsilon^*$, then MSEP = $m\sigma^2$ + PMSE. In discussing Dempster et al. (1977), Smith (1977) commented that the use of the PMSE criterion obscures many of the features of interest in the more general prediction

problem and may lead one to conclude falsely that there is less potential for improvement over OLS in the prediction context. He further commented that it is the relationship between the diagonal elements of $Z^{*}'Z^{*}$ and the corresponding eigenvalues of Z'Z which determines the potential for improvement in MSEP. There is a potential for improvement in MSEP when the large diagonal values of $Z^{*}'Z^{*}$ correspond to the small eigenvalues of Z'Z. When considering PMSE (instead of MSEP), the diagonal elements of $Z^{*}'Z^{*}$ are equal to the corresponding eigenvalues. Therefore, predicting at the usual model matrix explanatory values provides no insight regarding the greater potential for improvement.

A SAS program was used to generate random X'X matrices having different correlation structures. We considered three levels of multicollinearity for each fixed number of explanatory variables. That is, we generated the X'X matrices such that, for an r-variate X, we theoretically have weak, moderate, and strong multicollinearity. In particular, the explanatory variables have been generated in such a manner that certain groups of variables are correlated with each other.

We chose r = 6 and 15, and n = 50. The choices 6 and 15 were made so as to consider a relatively small and a reasonably large number of explanatory variables. Several authors have suggested that the improvement of the biased estimators over the OLS estimator is greater when r is large. However, Draper and Van Nostrand (1979) note that increasing r decreases the mean size of the estimated regression coefficients. They claim that the less significant a regression is, the greater the advantage of biased estimation over OLS estimation. We standardize the signal-to-noise ratio by dividing by r. For both choices of r, the degrees of freedom for s^2 are

larger than 30, which we believe is realistic. There is no evidence from the literature that sample size is an important factor in the relative performance of the biased estimators.

The explanatory variables were generated from the relationships:

$$X_{ij} = (1 - \tau^2)^{1/2} E_{ij} + \tau E_{is} ,$$

$$j = 1,2,\ldots,p \text{ and } i = 1,2,\ldots,50;$$

$$X_{ij} = (1 - \tau_*^2)^{1/2} E_{ij} + \tau_* E_{is} ,$$

$$j = p+1,p+2,\ldots,r \text{ and } i = 1,2,\ldots,50.$$

Here, $s = r+1$ and $E_{i1}, E_{i2},\ldots, E_{is}$ are independent N(0,1) variates. Thus, τ^2 is the theoretical correlation between any pair of random variables X_{i1}, $X_{i2}, \ldots, X_{ip}$, and $\tau\tau_*$ is the theoretical correlation between any random variable X_{ik} ($k = 1,2,\ldots,p$) and $X_{i\ell}$ ($\ell = p+1,\ldots,r$). Finally, τ_*^2 is the theoretical correlation between any pair of random variables $X_{i,p+1}$, $X_{i,p+2}, \ldots, X_{ir}$. The explanatory variables were standardized so that $X'X$ is in correlation form.

For both cases ($r = 6$ and 15), we introduce multicollinearity by considering two explanatory variables with a high correlation (i.e., $p = 2$). The values of τ and τ_* were taken to be such that $\tau^2 = .64$, $\approx .90$, and $\approx .98$, while $\tau_*^2 = .36$. See Table 1.

The eigenvalues and the normalized eigenvectors corresponding to the generated correlation matrices have been obtained. Table 1 also gives the maximum condition index, $\sqrt{\lambda_1/\lambda_r}$.

We chose β such that $\beta'\beta = \alpha'\alpha = q^2$, where $q^2 = .6$, 6, 60, 600, 6,000, and 60,000 for $r = 6$, and $q^2 = 1.5$, 15, 150, 1,500, and 15,000 for $r = 15$.

For each replication of the simulation, $\alpha = P'\beta$ was chosen such that α is the vector whose elements were generated at random from a N(0,1)

distribution and scaled such that α lies on the r-sphere having radius q. That is, r random numbers (e_i for i=1,2,...,r) were chosen from a N(0,1) distribution, and the individual regression coefficients were computed as $\alpha_i = qe_i / (\Sigma_{i=1}^{r} e_i^2)^{1/2}$. The true regression coefficients were generated at random for each replication.

In considering a value (or range of values) for σ^2, it is sufficient to assure a wide range of the signal-to-noise ratio (SNR = $\alpha'\alpha/\sigma^2$). Therefore, without loss of generality, we chose $\sigma^2 = 1$. The SNR is closely related to the expected F ratio and hence the significance of the regression.

Let $F_{r,n-r-1}(\nu)$ denote a noncentral F with degrees of freedom r, n-r-1, and noncentrality parameter $\nu = (1/2)\beta'X'X\beta$. If β is generated at random for each replication of the simulation from a symmetric (about zero) distribution, then Hoerl, Schuenemeyer, and Hoerl (1986) point out that E(F) = (1 + SNR/r)(n-r-1)/(n-r-3). We have chosen $q^2 = \beta'\beta$ such that the expected F ratios have a wide range of significance. The SNR has been standardized such that the expected F ratios are approximately equal for r = 6 and 15 (see Table 2).

For each simulation replication and for each of the 36 (X'X, β) simulation configurations, we generated $\hat{\alpha}$ from a $N(\alpha, A^{-1})$ distribution. The chi-square distribution with (50-r-1) degrees of freedom was used to generate the estimated variance.

Two sets of future observations were chosen for each of the cases (X'X matrices) as follows:

i) Choose r future observations equal to the r eigenvectors corresponding to the given X'X matrix.

ii) Choose r future observations at random in the same manner that the original data were generated and center and scale using the original X matrix.

We are interested both in predicting the value of one future observation at a time and simultaneously predicting more than one future observation.

For each simulation configuration, we generated 300 regressions. In other words, for a given $\mathbf{X'X}$ and $\boldsymbol{\beta}$, an α vector was generated for each of the 300 replications, and then $\hat{\alpha}$ was generated using the normal distribution with mean α.

Several calculations were made for each replication. All summary statistics were calculated for each single prediction and for multiple predictions associated with each predictor. That is, the set of 300 random numbers changed according to a given $\mathbf{X'X}$ and $\boldsymbol{\beta}$.

For each simulation configuration, and for each single prediction and for multiple predictions associated with each predictor, the estimated mean squared error of prediction and its estimated standard error were calculated. For example, the MSEP for the OLS predictor was estimated by $\hat{MSEP}(\hat{y}^*) = \Sigma_{\ell=1}^{300} SEP(\hat{y}^*)_\ell/300$, where $SEP(\hat{y}^*)_\ell = (\hat{y}^* - E(y^*))_\ell^2$ is the squared error for the ℓ-th replication for a given simulation configuration and prediction situation (single or multiple). The standard error of $\hat{MSEP}(\hat{y}^*)$ was then estimated as

$$s.e.\ \hat{MSEP}(\hat{y}^*) = \left[\frac{\Sigma_{\ell=1}^{300} SEP(\hat{y}^*)_\ell^2 - 300(MSEP(\hat{y}^*))^2}{300(299)} \right]^{1/2} .$$

The terms $m\sigma^2$ and $m\sigma^2/n$ (due to ϵ^* and the estimated intercept) in the

MSEP of a predictor are not included in the computation of the estimated MSEP.

The estimated bias and its standard error for the OLS predictor were calculated and used to check the accuracy of the simulation. Also, the average F statistic, $\bar{F} = \Sigma_{\ell=1}^{300} (\hat{\alpha}'\mathbf{A}\hat{\alpha})_{\ell}/(300(n-r-1))$, was calculated from the unbiased regression for each of the simulation configurations.

5. RESULTS OF THE SIMULATION STUDY

5.1 Accuracy of the Simulation Study

The numerical simulation was done on an IBM 3081 computer using SAS (Version 5.16). The normal deviates were generated using the RANNOR function (the Box-Muller transformation of uniform deviates). The chi-square deviates were generated using the RANGAM function.

For each single prediction from each simulation configuration, the estimated bias and its estimated standard error were calculated for the OLS predictor. Because the OLS predictor $\hat{y}^*$ is an unbiased predictor for $y^* = X^*\beta + \epsilon^*$, we would expect its estimated bias to be about zero. The usual z-statistics were calculated and used to test the hypothesis that the bias is zero. Less than five percent of the z-statistics were significant at the .05 level of significance.

The average F statistics and the corresponding standard errors were calculated for each simulation configuration. For all simulation configurations, the average F statistics were within two standard errors of the corresponding expected F statistics given in Table 2.

In this respect we believe that our simulation results are accurate.

5.2 Predicting in the Direction of the Eigenvectors of X'X

5.2.1 Predicting One Future Observation at a Time

If the response is predicted at x^*, where x^* is in the direction of one of the eigenvectors of X'X, then the GHK and OPK predictors computationally coincide.

When β (or α) is generated at random for each replication of the simulation study, we have, in a sense, averaged over many possible regression situations. Figures 1 through 6 illustrate the following results for r = 6, x^* equal to P_1, P_3, and P_6, and for high versus low levels of multicollinearity.

1. As the signal-to-noise ratio increases, the bias of the predictors tends to zero; hence, the estimated MSEP of the biased predictors converges to the estimated MSEP of the OLS predictor. In general, for small signal-to-noise ratios (SNR = .6, 6, 1.5 and 15), the OPK (GHK), OHK and OLW predictors offer substantial improvement over the OLS predictor.

2. As the direction of x^* changes from P_1 to P_r, through the r eigenvectors, the potential improvement of the biased predictors increases. When x^* is in the direction of the eigenvectors corresponding to the smallest eigenvalues, and SNR is small, the OR predictors, OHK and OLW offer the greatest potential improvement over the OLS predictor. However, when x^* is in the direction of the eigenvectors corresponding to the largest eigenvalues, and SNR is small, the OPK (GHK) predictor often offers greater potential improvement over the OLS predictor than do the OHK and OLW predictors.

3. For any fixed value of small to moderate signal-to-noise ratios (SNR = .6 to 600 and 1.5 to 1500), as the level of multicollinearity

increases, the performance of the OPK (GHK), OHK and OLW predictors improves with respect to the OLS predictor when $x^* = P_r$. This additional improvement (due to the increase in multicollinearity) in the performance of the biased predictors is the greatest for SNR = 60 and 150. For SNR = 6,000, the performance of the OLW predictor worsens as the level of multicollinearity increases.

4. When x^* is orthogonal to P_r, the potential improvement of the biased predictors is not affected by the level of multicollinearity (except for the OHK predictor). At times, the potential improvement of the OHK predictor with respect to the OLS predictor slightly decreases as the level of multicollinearity increases.

5. General statements regarding the potential improvement of the biased predictors with respect to the OLS predictor as the number of explanatory variables increases, cannot be made. The increase or decrease in potential improvement of the biased predictors depends on the direction of x^*. When x^* is in the direction of the eigenvectors corresponding to the larger eigenvalues (excluding λ_1, which is much larger than the other eigenvalues), an increase in r does not affect the performance of the OPK (GHK), OHK and OLW predictors. When x^* is in the direction of the eigenvectors corresponding to the smallest eigenvalues, an increase in r improves the performance of all of the biased predictors (except for the GPK predictor) relative to the OLS predictor.

6. In some cases with large signal-to-noise ratios, the estimated MSEP of the OPK (GHK), and OLW predictors are slightly larger than the estimated MSEP of the OLS predictor.

5.2.2 Predicting r Future Observations at a Time

When $X^{*}{}'X^{*} = I$, the estimated MSEP = MSE + $r\sigma^2$. Therefore, the GPK predictor reduces to the GHK predictor, and the OPK predictor reduces to the OHK predictor with $\hat{c}_{OHK} = rs^2/\hat{\alpha}'\hat{\alpha}$. Any improvement of the OHK predictor as compared to the OPK predictor is due to the iterative selection of $\hat{c}_{OHK}$.

Based on MSE, the GPK (GHK) predictor performs well when compared to the OLS predictor. However, it does not perform as well as the OPK, OHK, and OLW predictors. These results agree with those of Lawless and Wang (1976). In Figures 7 and 8, we compare the predictors when $X^{*} = P$ for low and high levels of multicollinearity (r = 6).

When r is increased to 15, we observe that the performance of the OR predictors improves. The increase does not affect the performance of the GR predictors.

5.3 Predicting in a Randomly Chosen Direction

5.3.1 Predicting One Future Observation at a Time

Recall that r future observations were chosen at random in the same manner as the original data were generated. These vectors have not been scaled to have length equal to one. However, they have been centered and scaled by the original X matrix.

Considering x^{*} as a linear combination of the eigenvectors, the performance of the biased predictors depends on the coefficients of the linear combination. In general, when x^{*} is oriented towards the eigenvectors corresponding to the smallest eigenvalues, the potential improvement of the biased predictors is increased. However, since x^{*} is not necessarily parallel to one of the r eigenvectors, the GHK and OPK

predictors are no longer computationally identical. In this case, the performance of the GHK and OPK predictors is similar for small signal-to-noise ratios. For moderate to large signal-to-noise ratios, the performance of the OPK predictor can be very poor, particularly for a high level of multicollinearity.

In Figures 9 and 10, we give two examples where x^* is randomly generated. For $x^* = x_1^*$ in Figure 9, the coefficients $P'x^*$ are equal to $(.03\ -.28\ -.64\ -.56\ -.31\ -.31)'$ and the level of multicollinearity is low. Note that x_1^* is an "average" of P_2 through P_6. The relative performance of the GPK, GHK, OHK, and OLW predictors is similar to that observed in Figures 2 and 3. The performance of the OPK predictor is poor for large signal-to-noise ratios. Similar results were observed for the five other future observations randomly generated in our simulation. Hence, those cases are not included here. For $x^* = x_2^*$ in Figure 10, the coefficients are equal to $(-.11\ -.02\ -.36\ .60\ .64\ .29)'$ and the level of multicollinearity is high. The orientation of x_2^* is more towards the eigenvectors corresponding to the smaller eigenvalues. The relative performance of the GPK, GHK, OHK and OLW is similar to that observed in Figure 6, thus showing an increase in performance due to an increased level of multicollinearity. The performance of the OPK predictor when $x^* = x_2^*$ as compared to $x^* = P_6$ is better for small signal-to-noise ratios and much worse for large signal-to-noise ratios.

5.3.2 Predicting r Future Observations at a Time

Recall that the estimated MSEP for simultaneous multiple predictions for the GHK, OHK, OLW and OLS is the sum of the MSEPs for single predictions. Therefore, the remarks presented for single predictions

extend "in an average sense" to these four predictors.

For simultaneous prediction of X^* randomly chosen in our simulation study, Figures 11 and 12 give the relative performance of the biased predictors with respect to the OLS predictor. The performance of the GPK and GHK predictors is similar. For small signal-to-noise ratios, the OPK and OLW predictors out perform the others. For moderate to large signal-to-noise ratios, the performance of the OPK predictor can be slightly worse than that of the OLS predictor. In our examples, the performance of the OPK predictor is significantly improved when simultaneously predicting several future observations.

Finally, we will discuss these results and make recommendations for using biased predictors.

6. DISCUSSION AND RECOMMENDATIONS

The performance of the biased predictors relative to the OLS predictor depends on several important "parameters". It strongly depends on the direction(s) in which we wish to make our predictions. It also depends on the magnitude of the signal-to-noise ratio, the level of multicollinearity, and the number of explanatory variables.

The results given in Section 5 indicate that, under certain circumstances, the traditional biased predictors offer a substantial improvement in estimated MSEP over the OLS predictor, while the performance of the proposed biased predictors can be ruled out as alternatives to the OLS predictor.

In general, the traditional biased predictors are known to offer good

estimation of the true regression coefficient vector, but such estimation does not depend on X^*. Good estimation of the true regression coefficient vector may not necessarily be sufficient to insure good prediction. It is also possible to have rather poor estimation and yet have good prediction.

Unfortunately, the predictors based on optimal prediction (minimum MSEP), for which the optimal K values depend on X^*, do not offer improvement over the traditional biased predictors even when considering prediction at multiple x^*'s. This is particularly true when the signal-to-noise ratio is small.

For this reason, the GPK and OPK predictors can essentially be ruled out as alternatives to the OLS predictor. However, the GPK predictor offers the advantage that its estimated MSEP is essentially always less than that of the OLS predictor. The OPK predictor performs poorly for large signal-to-noise ratios when X^* is chosen at random in the same manner as the original data were generated. When x^* is equal to an eigenvector of X'X, the OPK predictor is equal to the GHK predictor. When $X^{*\prime}X^* = I$, the OPK predictor is equal to the uniterated OHK predictor, which is inferior to the iterated OHK predictor.

The performance of the remaining traditional predictors can be substantially better than the OLS predictor. In general, the OHK and OLW predictors perform the best with respect to the OLS predictor for small signal-to-noise ratios, but can have large estimated MSEP for moderate to large signal-to-noise ratios under particular circumstances. The performance of the GHK predictor is never much worse than the OLS predictor for any signal-to-noise ratio and can be substantially better than the OLS predictor for moderate and large signal-to-noise ratios.

It was found that the performance of the predictors does not necessarily improve as the level of multicollinearity or the number of explanatory variables increases. However, the relative performance depends on the direction in which the prediction is made.

Finally, we make the following recommendations. If the noise (σ^2) is (intuitively) too large compared to the signal ($\boldsymbol{\beta}'\boldsymbol{\beta}$), then none of the predictors, including the OLS predictor, will give good predictions. Otherwise, estimate the signal-to-noise ratio as $\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}/s^2$. [Note that $E(\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}/s^2) = (SNR + \Sigma_{i=1}^{r} 1/\lambda_i)(n-r-1)/(n-r-3)$, and hence one may also use $[\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}\ (n-r-3)/(s^2(n-r-1))] - \Sigma_{i=1}^{r} 1/\lambda_i$ as an estimate of the signal-to-noise ratio.] If the estimated signal-to-noise ratio (divided by r) is large (i.e., greater than 100), then use the OLS predictor. If the estimated signal-to-noise ratio is moderate, say between 10 and 100, then the OHK predictor may offer substantial improvement. If the estimated signal-to-noise ratio is small, say less than 10, then the OLW predictor offers the most improvement over the OLS predictor. The OLW predictor performs better than the OHK predictor for small signal-to-noise ratios, but is not necessarily better for moderate signal-to-noise ratios.

Other authors, including Gibbons (1981) and Wichern and Churchill (1978), have compared ridge estimators when $\boldsymbol{\beta}$ is fixed and in the direction of the eigenvectors corresponding to the largest and smallest eigenvalues. For fixed $\boldsymbol{\beta}$, results involving the predictors discussed here have been obtained by Pantula (1987) in a prediction context. In addition, Pantula included comparisons that were made using principal components and shrunken least squares predictors for both fixed and random $\boldsymbol{\beta}$.

APPENDIX

In this Appendix we present the proofs of Results 1-4.

Proof of Result 1: Let $\mathbf{\Lambda} = \text{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_r\}$ be the eigenvalues of $\mathbf{X'X}$ and let $\mathbf{K} = \text{diag}\{k_1, k_2, \ldots, k_r\}$. Then we can write $\text{MSEP}(\tilde{\mathbf{y}}_K^*)$, given in (2.2), as

$$\text{MSEP}(\tilde{\mathbf{y}}_K^*) = \sigma^2[m + \sum_{i=1}^{r} f_i k_i^2] + \sum_{\ell=1}^{m} [\sum_{i=1}^{r} z_{\ell i}^* \alpha_i (1 - k_i)]^2,$$

where $f_i = \sum_{\ell=1}^{m} z_{\ell i}^{*2}/\lambda_i$.

Setting the first derivative with respect to k_i to zero, we obtain

$$\sigma^2 f_i \tilde{k}_i + \sum_{j=1}^{r} \sum_{\ell=1}^{m} z_{\ell j}^* z_{\ell i}^* \alpha_i \alpha_j \tilde{k}_j = \sum_{\ell=1}^{m} \sum_{j=1}^{r} z_{\ell i}^* z_{\ell j}^* \alpha_i \alpha_j \quad ,$$

for $i = 1,2,\ldots,r$. Expressing the above equations in the matrix form we get $\mathbf{A}\tilde{\mathbf{k}} = \mathbf{b}$, where $\mathbf{A}$ and $\mathbf{b}$ are defined in (2.3). It is easy to see that $\mathbf{A}$ is p.d. and therefore, $\mathbf{A}^{-1}$ exists. Furthermore, the solution $\tilde{\mathbf{k}}$ corresponds to a minimum, since the second derivative of $\text{MSEP}(\tilde{\mathbf{y}}_K^*) = \mathbf{A}$ is p.d. []

Proof of Result 2: Note that the matrix $\mathbf{Z^{*'}Z^*}$ reduces to $(z^*z^{*'})_{ij} = ((z_i^* z_j^*))$. Then, the system of equations $\mathbf{A}\tilde{\mathbf{k}} = \mathbf{b}$ reduces to

$$z_i^{*2}(\alpha_i^2 + \sigma^2/\lambda_i)\tilde{k}_i + \sum_{\substack{j=1 \\ j\neq i}}^{r} z_i^* z_j^* \alpha_i \alpha_j \tilde{k}_j = \alpha_i z_i^* \sum_{j=1}^{r} z_j^* \alpha_j \quad ,$$

for $i=1,2,\ldots,r$. Therefore,

$$k_i = (\alpha_i \lambda_i \,/\, z_i^* \sigma^2)\, c \quad ,$$

where $c = \sum_{j=1}^{r} (1 - \tilde{k}_j) z_j^* \alpha_j$ is a constant for all i. Solving for c, we have

$$c = \sum_{j=1}^{r} z_j^* \alpha_j \,/\, (1 + \sum_{j=1}^{r} \alpha_j^2 \lambda_j \,/\, \sigma^2) \quad . \qquad []$$

Proof of Result 3: Since $\mathbf{X}'\mathbf{X} = \mathbf{I}$, we have $\mathbf{\Lambda} = \mathbf{I}$ and $\mathbf{K}_{OR} = (\mathbf{\Lambda} + c_{OR}\mathbf{I})^{-1}\mathbf{\Lambda} = d_{OR}\mathbf{I}$, where $d_{OR} = 1/(1 + c_{OR})$. Now taking the derivative of $\text{MSEP}(\tilde{y}^*_{OR})$ with respect to d_{OR} and setting it to zero, we obtain

$$\tilde{d}_{OR} = \alpha'Z^{*\prime}Z^*\alpha \; / \; [\sigma^2 \; \text{tr}(Z^{*\prime}Z^*) + \alpha'Z^{*\prime}Z^*\alpha] \; .$$

Therefore,

$$\tilde{c}_{OR} = \sigma^2 \; \text{tr} \; (Z^{*\prime}Z^*) \; / \; \alpha'Z^{*\prime}Z^*\alpha \; .$$

Since the second derivative of $\text{MSEP}(\tilde{y}^*_{OR})$ is positive, the solution corresponds to a minimum. []

Proof of Result 4:

Note that

$$\hat{y}^*_{GPK} = z^{*\prime}\hat{K}_{GPK}\,\hat{\alpha} = z^{*\prime} \; \text{diag} \; \{q_1, q_2, \ldots, q_r\} \; \hat{\alpha} \; \hat{y}^* ,$$

where $q_i = \hat{\alpha}_i \; \lambda_i \; / \; z_i^*(s^2 + \Sigma_{j=1}^r \hat{\alpha}_j^2 \; \lambda_j)$.

Therefore,

$$\hat{y}^*_{GPK} = z^{*\prime} \; \{\tilde{q}_1, \tilde{q}_2, \ldots, \tilde{q}_r\}' \; \hat{y}^*$$

$$= (\Sigma_{j=1}^r \hat{\alpha}_j^2 \; \lambda_j) \; \hat{y}^* \; / \; (s^2 + \Sigma_{j=1}^r \hat{\alpha}_j^2 \; \lambda_j) \; ,$$

where $\tilde{q}_i = q_i\hat{\alpha}_i$.

Since, $0 \leq \Sigma_{j=1}^r \hat{\alpha}_j^2 \; \lambda_j \; / \; (s^2 + \Sigma_{j=1}^r \hat{\alpha}^2 \; \lambda_j) < 1$, we have that $\hat{y}^*_{GPK}$ is a shrunken value of $\hat{y}^*$. []

REFERENCES

Copas, J. B. (1983). Regression, prediction and shrinkage (with discussion). Journal of the Royal Statistical Society, Series B **45**, 311-354.

Dempster, A. P., Schatzoff, M. & Wermuth, N. (1977). A simulation study of alternatives to ordinary least squares (with discussion). Journal of the American Statistical Association, **72**, 77-103.

Draper, N. R. & Van Nostrand, R. C. (1979). Ridge regression and James-Stein estimation: Review and comments. Technometrics **21**, 451-466.

Gibbons, D. G. (1981). A Simulation Study of Some Ridge Estimators. Journal of the American Statistical Association, **76**, 131-139.

Hoerl, A. E. & Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonol problems. Technometrics **12**, 55-67.

Hoerl, A. E. & Kennard, R. W. (1976). Ridge regression: Iterative estimation of the biasing parameter. Communications in Statistics **A5(1)**, 77-88.

Hoerl, A. E., Kennard, R. W. & Baldwin, K. F. (1975). Ridge regression: Some simulations. Communications in Statistics **4(2)**, 105-123.

Hoerl, R. W., Schuenemeyer, J. H. & Hoerl, A. E. (1986). A simulation of biased estimation and subset selection regression techniques. Technometrics **28**, 369-380.

Lawless, J. F. & Wang, P. (1976). A simulation study of ridge and other regression estimators. Communications in Statistics **A5(4)**, 307-323.

Obenchain, R. L. (1978). Good and optimal ridge estimators. The Annals of Statistics **6**, 1111-1121.

Pantula, J. F. (1987). "Optimal Prediction in Linear Regression Analysis," unpublished Ph.D. dissertation, Institute of Statistics Mimeo Series No. 1837T, University of North Carolina at Chapel Hill, Dept. of Biostatistics.

Smith, A. F. M. (1977). Discussion of Dempster, A. P., Schatzoff, M. & Wermuth, N. A simulation study of alternatives to ordinary least squares. Journal of the American Statistical Association **72**, 77-103.

Vinod, H. D. (1976). Simulation and extension of a minimum mean squared error estimator in comparison with Stein's. Technometrics **18**, 491-496.

Wichern, D. W. & Churchill, G. A. (1978). A comparison of ridge estimators. Technometrics **20**, 301-311.

Table 1. Theoretical Correlations Between Pairs of Explanatory Variables for the Six Cases Considered in Simulation Study (p = 2)

	r	τ^2	$\tau\tau_*$	τ_*^2	Condition Index $\sqrt{\lambda_1 / \lambda_r}$
Case 1	6	.6400	.4800	.3600	$\sqrt{3.0663/.2828} = 3.29$
Case 2	6	.9025	.5700	.3600	$\sqrt{3.2909/.1413} = 4.83$
Case 3	6	.9801	.5940	.3600	$\sqrt{3.8764/.0146} = 16.28$
Case 4	15	.6400	.4800	.3600	$\sqrt{6.7503/.1191} = 7.53$
Case 5	15	.9025	.5700	.3600	$\sqrt{7.2039/.0440} = 12.80$
Case 6	15	.9801	.5940	.3600	$\sqrt{6.8332/.0199} = 18.52$

Table 2. Expected F Ratios for r = 6 and 15 in Simulation Study

r = 6			r = 15	
SNR/r	SNR	$E(F_{6,43}(\nu))^*$	SNR	$E(F_{15,34}(\nu))^*$
.1	.6	1.15	1.5	1.16
1	6	2.10	15	2.13
10	60	11.54	150	11.69
100	600	105.93	1,500	107.31
1,000	6,000	1,049.80	15,000	1,063.56
10,000	60,000	10,488.85	150,000	10,626.06

*n = 50

Figure 1. Ratio of the Empirical MSEP of each Biased Predictor to the MSEP of the OLS Predictor as a Function of the Signal-to-Noise Ratio, for $x^* = P_1$ and for Low Multicollinearity.

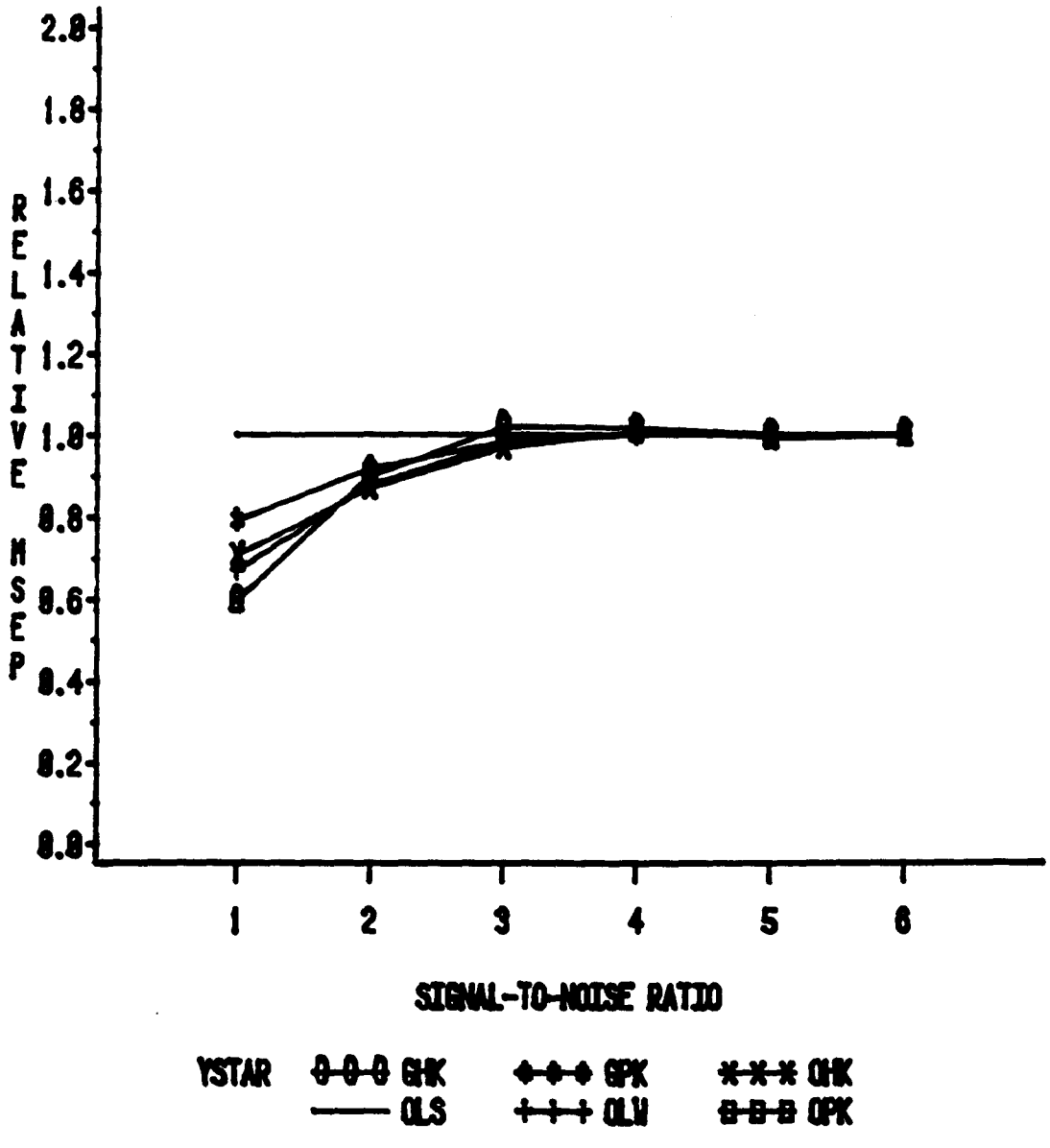

Figure 2. Ratio of the Empirical MSEP of each Biased Predictor to the MSEP of the OLS Predictor as a Function of the Signal-to-Noise Ratio, for $x^* = P_3$ and for Low Multicollinearity.

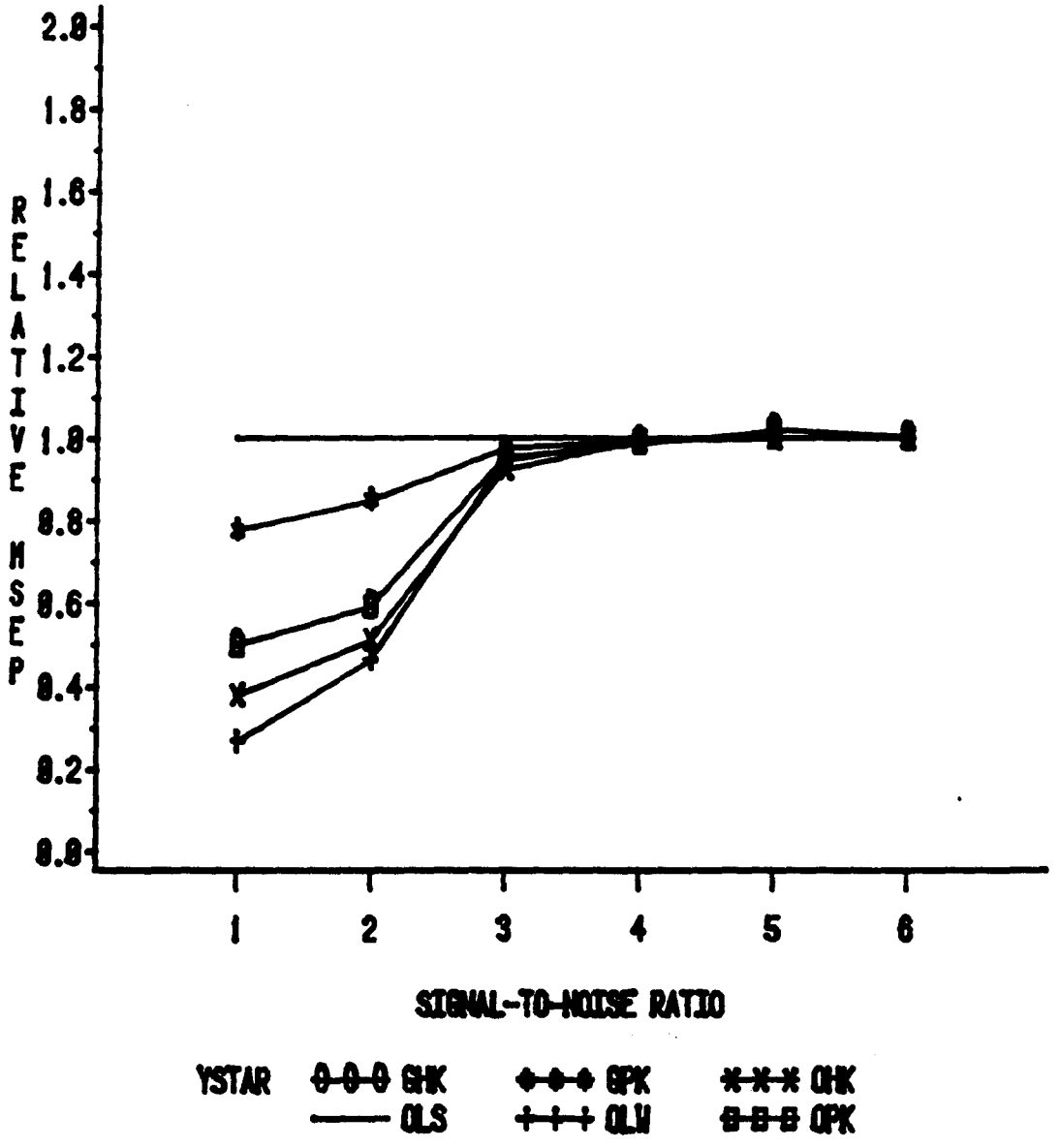

Figure 3. Ratio of the Empirical MSEP of each Biased Predictor to the MSEP of the OLS Predictor as a Function of the Signal-to-Noise Ratio, for $x^* = P_6$ and for Low Multicollinearity.

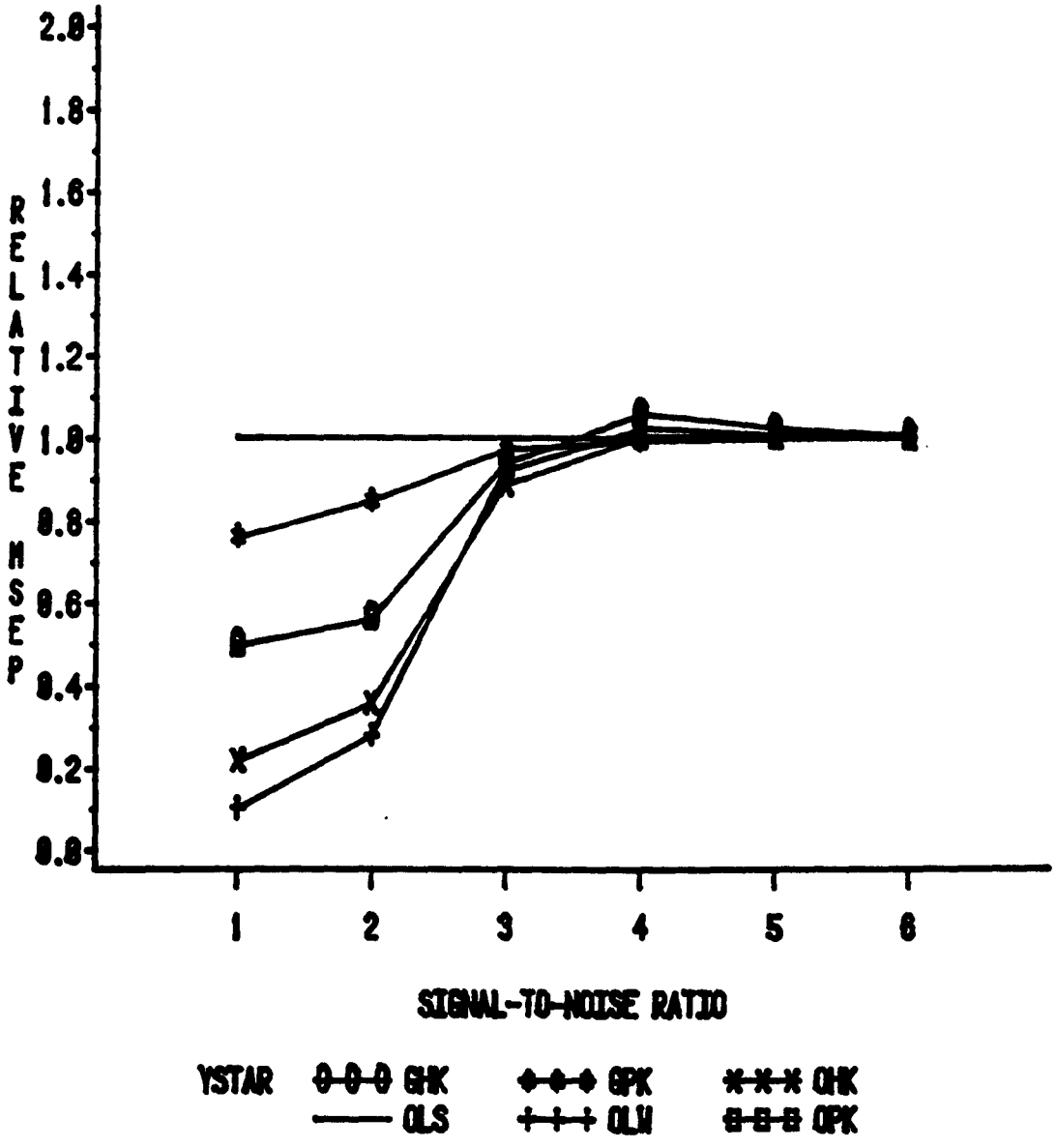

Figure 4. Ratio of the Empirical MSEP of each Biased Predictor to the MSEP of the OLS Predictor as a Function of the Signal-to-Noise Ratio, for $x^* = P_1$ and for High Multicollinearity.

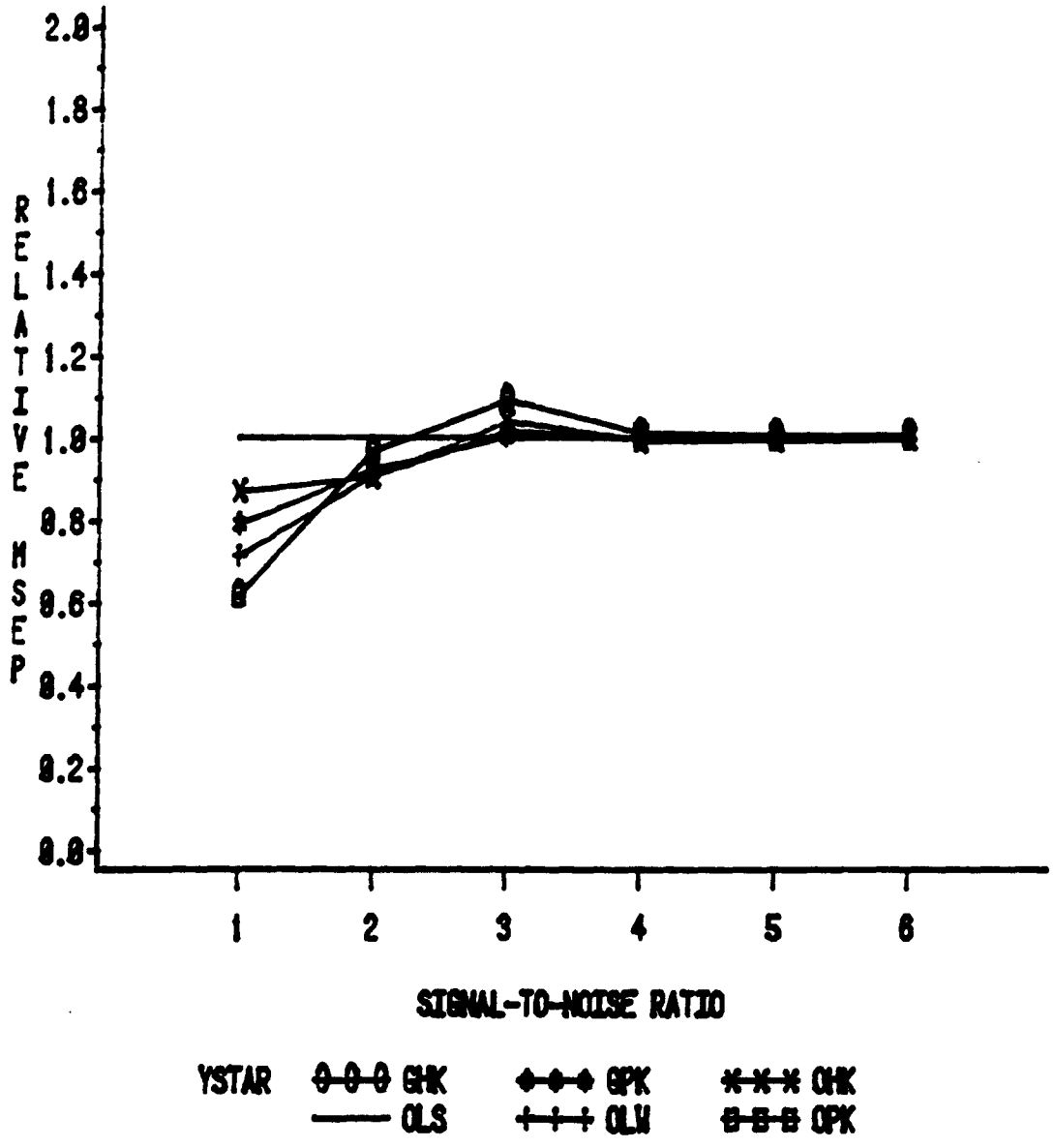

Figure 5. Ratio of the Empirical MSEP of each Biased Predictor to the MSEP of the OLS Predictor as a Function of the Signal-to-Noise Ratio, for $x^* = P_3$ and for High Multicollinearity.

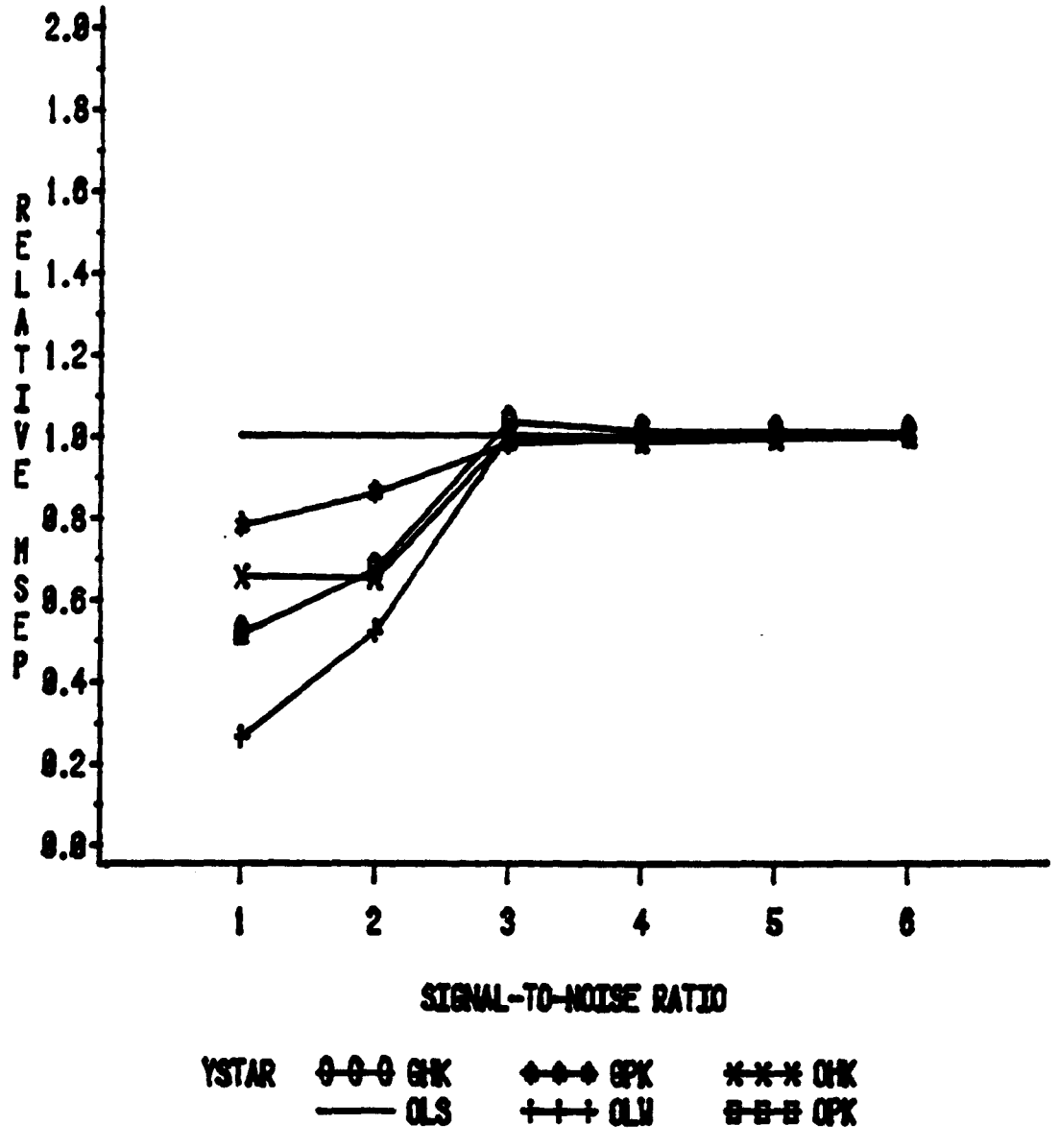

Figure 6. Ratio of the Empirical MSEP of each Biased Predictor to the MSEP of the OLS Predictor as a Function of the Signal-to-Noise Ratio, for $x^* = P_6$ and for High Multicollinearity.

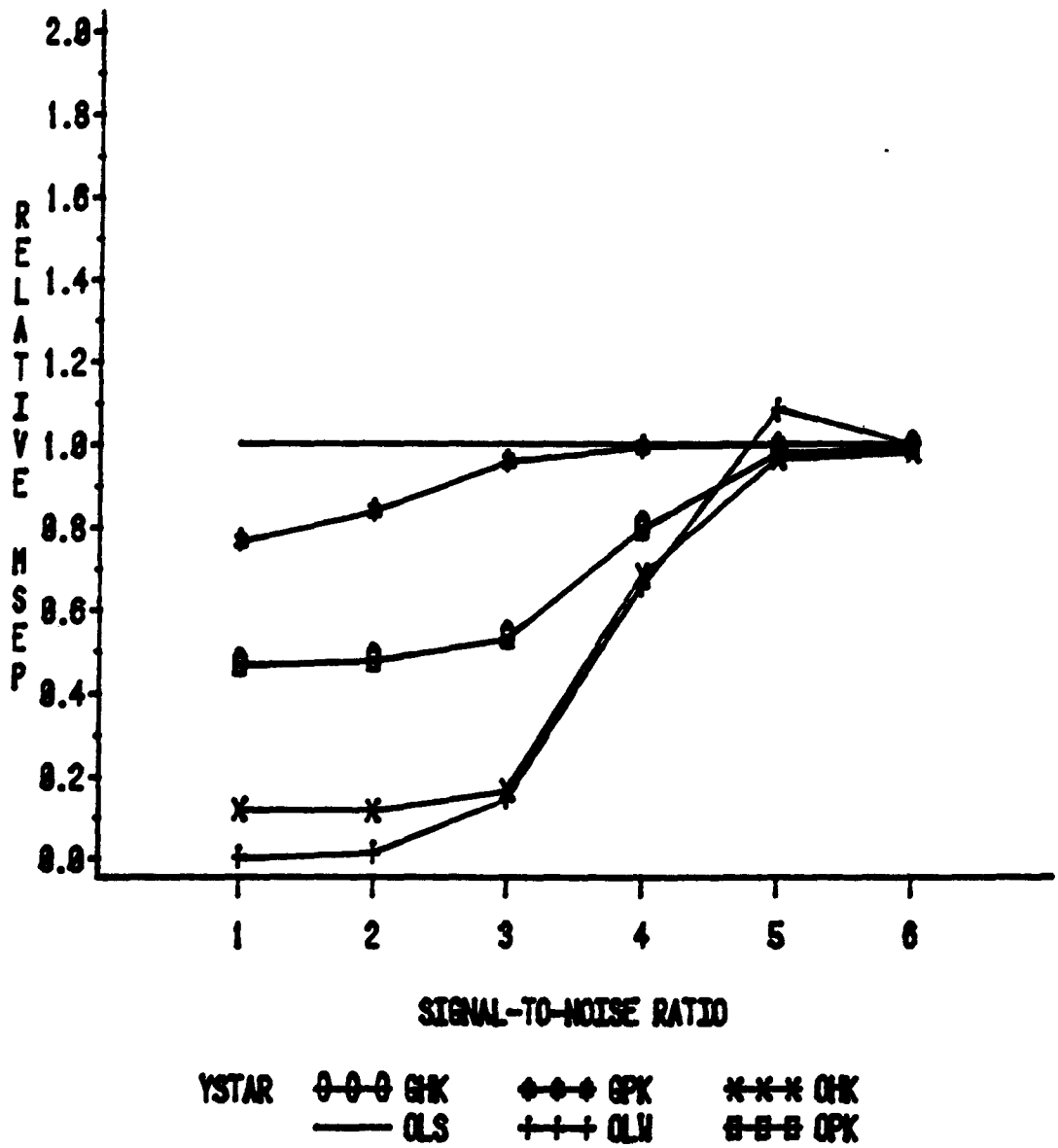

Figure 7. Ratio of the Empirical MSEP of each Biased Predictor to the MSEP of the OLS Predictor as a Function of the Signal-to-Noise Ratio, for $x^* = x_i^*$ and for Low Multicollinearity.

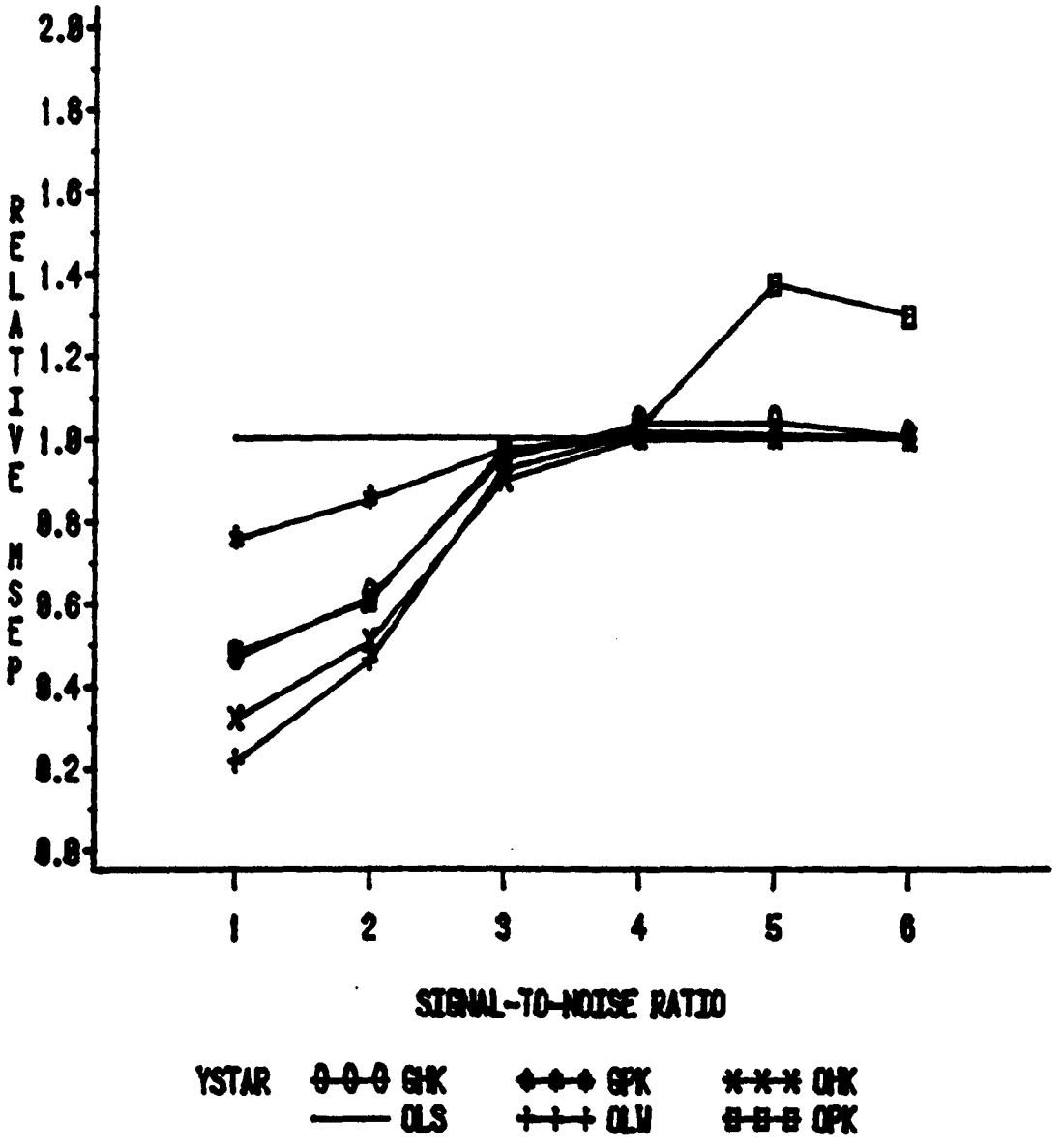

Figure 8. Ratio of the Empirical MSEP of each Biased Predictor to the MSEP of the OLS Predictor as a Function of the Signal-to-Noise Ratio, for $x^* = x_2^*$ and for High Multicollinearity.

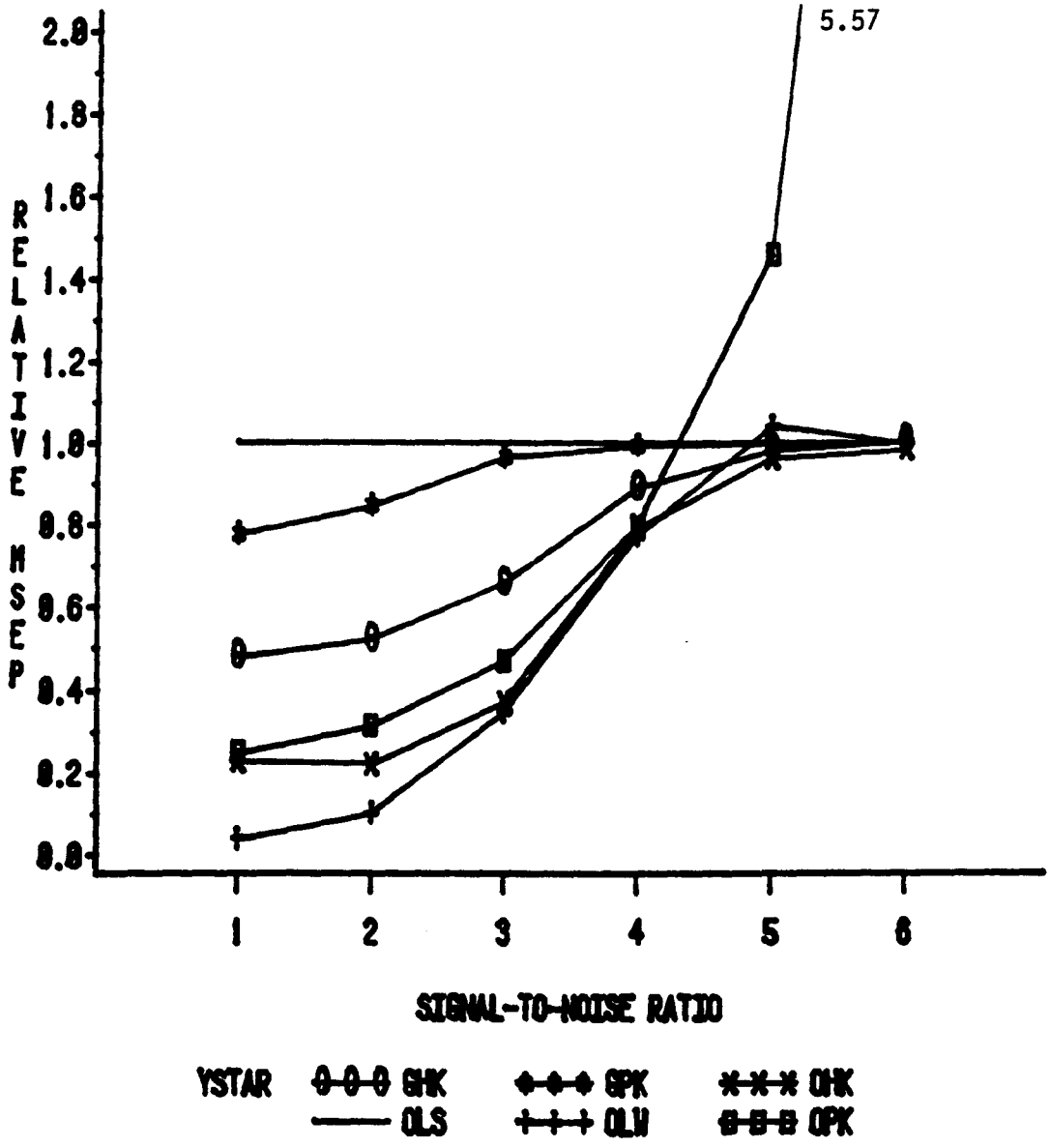

Figure 9. Ratio of the Empirical MSEP of each Biased Predictor to the MSEP of the OLS Predictor as a Function of the Signal-to-Noise Ratio, for $X^* = P$ and for Low Multicollinearity.

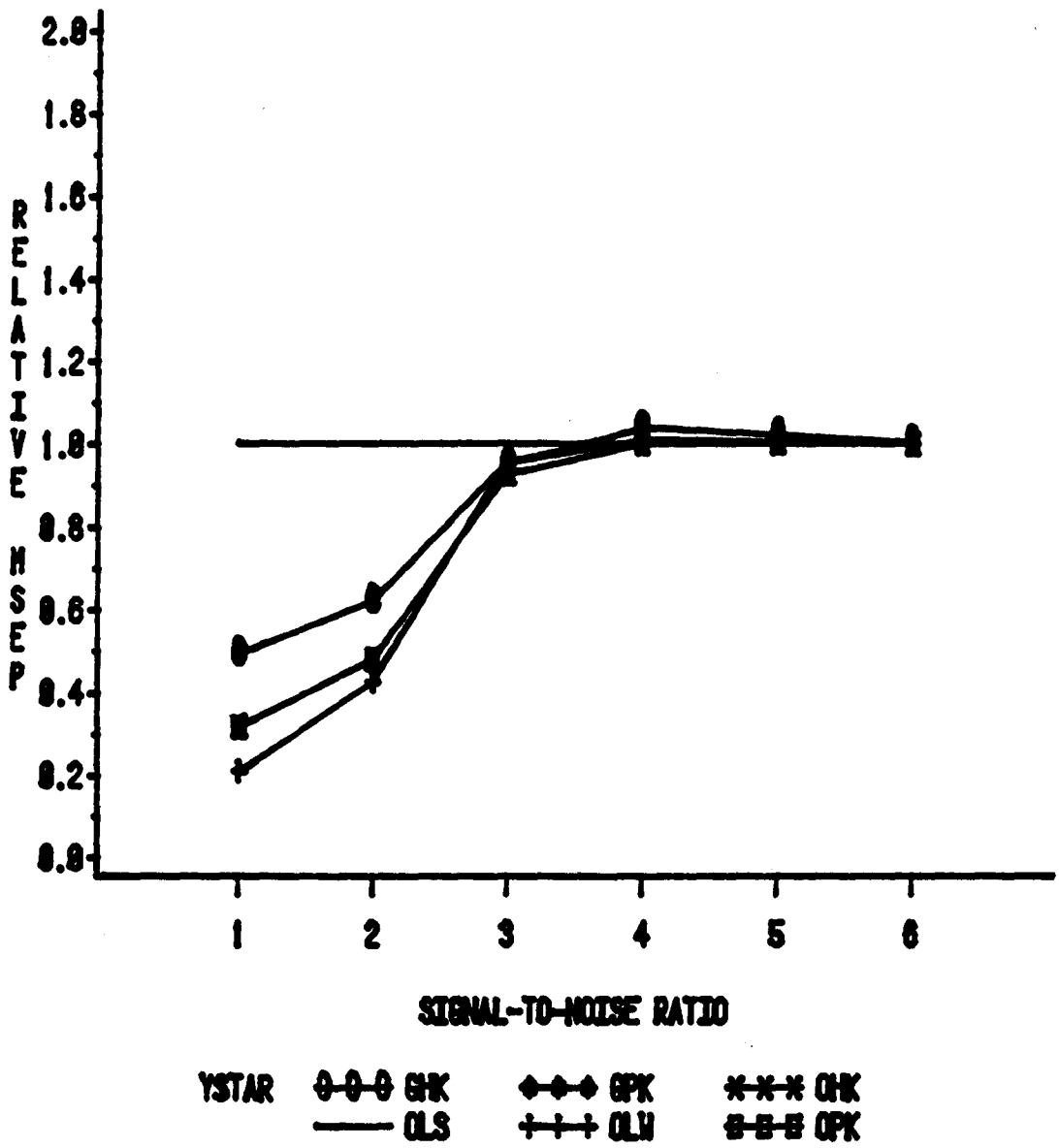

Figure 10. Ratio of the Empirical MSEP of each Biased Predictor to the MSEP of the OLS Predictor as a Function of the Signal-to-Noise Ratio, for $X^* = P$ and for High Multicollinearity.

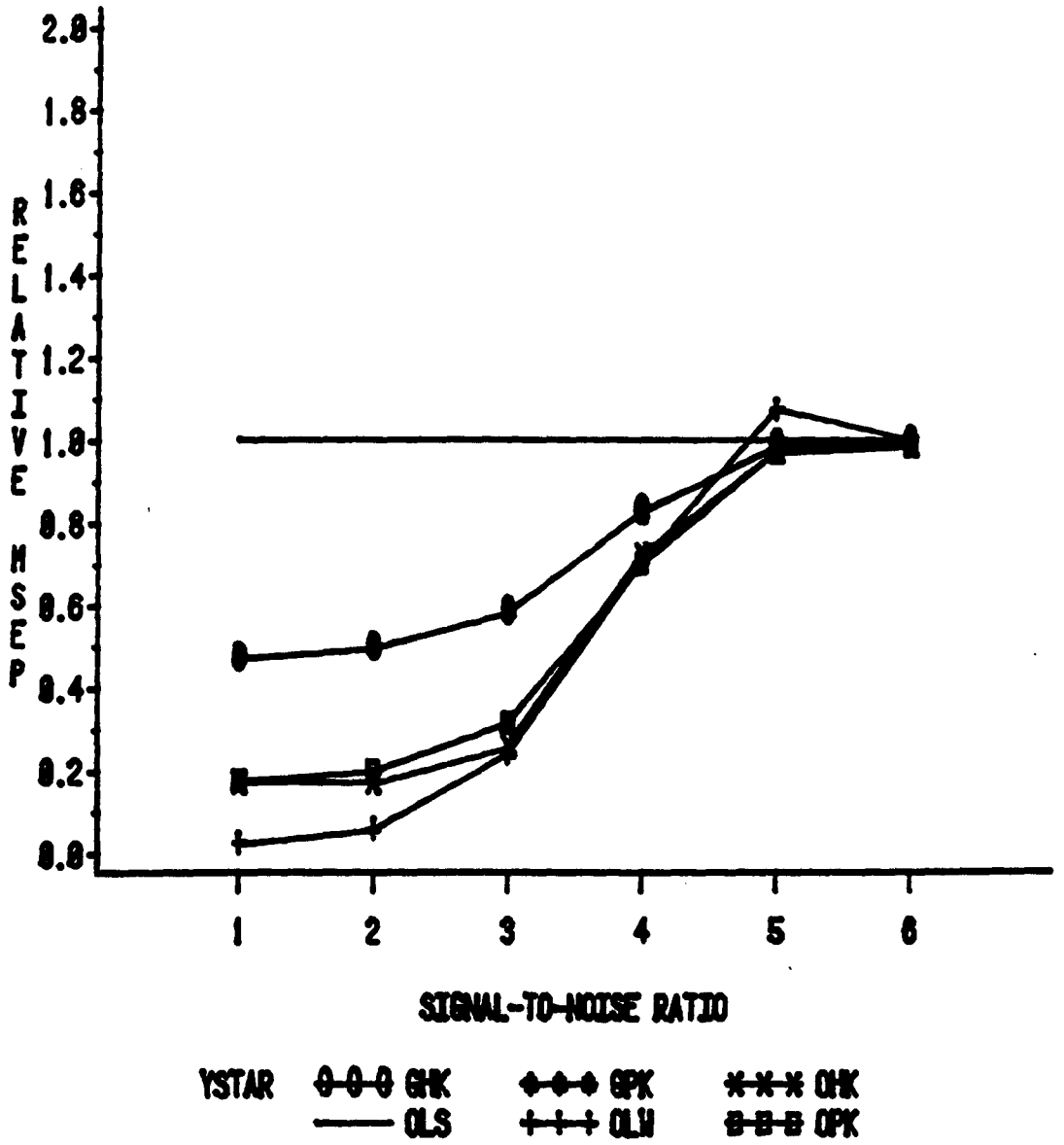

Figure 11. Ratio of the Empirical MSEP of each Biased Predictor to the MSEP of the OLS Predictor as a Function of the Signal-to-Noise Ratio, for Randomly Generated X^* and for Low Multicollinearity.

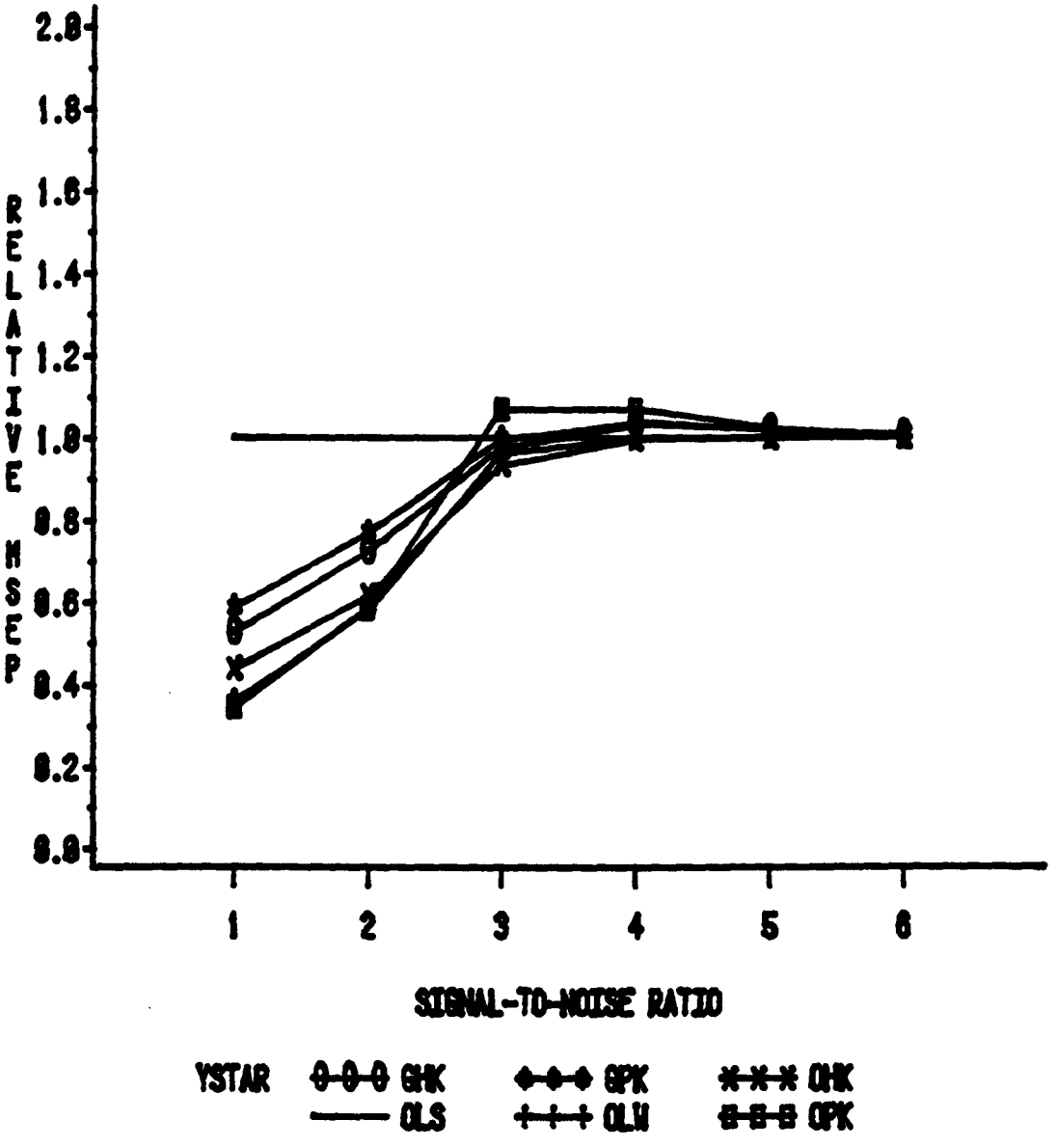

Figure 12. Ratio of the Empirical MSEP of each Biased Predictor to the MSEP of the OLS Predictor as a Function of the Signal-to-Noise Ratio, for Randomly Generated X^* and for High Multicollinearity.