The Annals of Statistics
1994, Vol. 22, No. 4, 1947–1975

THE RISK INFLATION CRITERION FOR MULTIPLE REGRESSION

BY DEAN P. FOSTER AND EDWARD I. GEORGE

University of Pennsylvania and University of Texas, Austin

A new criterion is proposed for the evaluation of variable selection procedures in multiple regression. This criterion, which we call the risk inflation, is based on an adjustment to the risk. Essentially, the risk inflation is the maximum increase in risk due to selecting rather than knowing the "correct" predictors. A new variable selection procedure is obtained which, in the case of orthogonal predictors, substantially improves on AIC, C_p and BIC and is close to optimal. In contrast to AIC, C_p and BIC which use dimensionality penalties of 2, 2 and $\log n$, respectively, this new procedure uses a penalty $2 \log p$, where p is the number of available predictors. For the case of nonorthogonal predictors, bounds for the optimal penalty are obtained.

0. Introduction. Consider the problem where, based on the observation of a dependent variable Y and a large set of potential predictors $X_1, \ldots, X_p$, one would like to build the "best" multiple regression model. More precisely, one would like to find and fit the "best" linear regression model of the form $Y = X_1^* \beta_1^* + \cdots + X_q^* \beta_q^* + \varepsilon$, where $X_1^*, \ldots, X_q^*$ is a "selected" subset of $X_1, \ldots, X_p$. A popular strategy is first to use a criterion based on the data to select $X_1^*, \ldots, X_q^*$, and second to estimate the coefficients $\beta_1^*, \ldots, \beta_q^*$ by "least squares." We shall refer to such a two-stage strategy as a selection/estimation (s/e) procedure. [An s/e procedure was called "subset least squares" by Mallows (1973).] The vague and often unstated goal of such s/e procedures is to achieve a desirable trade-off between predictive or explanatory power and parsimony.

Variable selection procedures for choosing a desirable subset of predictors abound. A partial list of procedures motivated by a wide variety of criteria includes adjusted R^2 [Theil (1961)], FPE [Akaike (1970)], posterior odds [Zellner (1971)], A_p [Allen (1971)], C_p [Mallows (1973)], PRESS [Allen (1974)], AIC [Akaike (1974)], S_p [Hocking (1976)], BIC [Schwarz (1978)], $2 \log \log n$ [Hannan and Quinn (1979)], PMDL [Risannen (1986)] and FIC [Wei (1992)]. A comprehensive summary of variable selection procedures as well as an extensive bibliography can be found in the recent book by Miller (1990).

In this paper a new criterion is proposed for the evaluation of variable selection procedures in multiple regression. This criterion, which we call risk inflation, is the maximum possible increase in risk of the consequent s/e procedure due to selecting rather than knowing the "correct" predictors. The risk inflation is obtained as the ratio of risk of an s/e estimator to the risk of the ideal (but unavailable) s/e estimator which uses only the "correct" predictors. Although risk inflation may used with any risk definition, in this paper we focus

Received October 1990; revised November 1993.
AMS 1991 *subject classifications.* Primary 62C99; secondary 62J05, 62C20.
Key words and phrases. Decision theory, minimax, model selection, multiple regression, risk, variable selection.

on the special case of risk inflation under predictive risk. As opposed to using unadjusted predictive risk, which from a minimax perspective forces inclusion of all predictors, use of the risk inflation criterion favors variable selection.

For the case of orthogonal predictors, it is seen that compared to overall inclusion, the popular variable selection procedures AIC, C_p and BIC offer smaller risk inflation. Moreover, a new variable selection procedure is obtained which, again in the case of orthogonal predictors, substantially improves on AIC, C_p and BIC and is close to optimal. In contrast to AIC, C_p and BIC which use dimensionality penalties of 2, 2 and $\log n$, respectively, this new procedure uses a penalty $2 \log p$, where p is the number of available predictors. Unfortunately, in the case of nonorthogonal predictors, the optimal dimensionality penalty depends on the correlation structure of the predictors. Bounds for the optimal penalty are obtained for this case.

Section 1 formalizes the notion of the predictive risk of an s/e procedure. Section 2 defines and motivates the risk inflation criterion. Section 3 defines a general canonical variable selection procedure which, when σ^2 is known, includes the selection procedures AIC, C_p and BIC. Section 4 obtains the risk inflation of a variety of variable selection procedures and provides the optimal procedure when the predictors are orthogonal and σ^2 is known. Section 5 obtains general bounds for the risk inflation for case of nonorthogonal predictors. Section 6 treats the more realistic situation where σ^2 is unknown. It is seen that all of the previous results are extendable to this situation. Section 7 concludes with the definition of a new variable selection procedure for the general case, whose risk inflation properties in many situations will be close to optimal. Finally, to improve the readability of the main text, the proofs of the main theorems, as well as some necessary lemmas, have been placed in the Appendix. However, we should point out that these proofs and lemmas may be of independent interest. For example, the proof of Lemma A.2 is based on using data augmentation in order to obtain a sufficiency reduction. Also, some of the lemmas highlight aspects of predictive risk not mentioned in the text.

1. The risk of an s/e procedure. Consider the following canonical decision-theoretic setup for fitting a multiple regression model. Let

$$Y = X_1\beta_1 + \cdots + X_p\beta_p + \varepsilon = X\beta + \varepsilon, \tag{1.1}$$

where Y is $n \times 1$, $X = [X_1, \ldots, X_p]$ is $n \times p$, $\beta = (\beta_1, \ldots, \beta_p)'$ is $p \times 1$, and $\varepsilon \sim N_n(0, \sigma^2 I)$. Also $X_1 \equiv (1, \ldots, 1)'$ so that β_1 is the intercept term. In this setting an estimator $\hat{\beta}$ of β is evaluated by its risk $R(\beta, \hat{\beta})$ which will be the expected loss for an appropriate loss function. Throughout this paper we focus on using the predictive risk,

$$R(\beta, \hat{\beta}) = E_\beta |X\hat{\beta} - X\beta|^2. \tag{1.2}$$

This risk may be considered to be the expected squared error of prediction if X is representative of future prediction values. [Note that we are treating X as fixed in the sense of Thompson (1978a, b).] Note also that under (1.2) the

estimation problem is invariant under location and scale transformations. The predictive risk has also been used in the context of evaluating variable selection procedures by Mallows (1973), Shibata (1981) and Miller (1990). However, it should be mentioned that our development of the risk inflation criterion in Section 2 can be carried out for arbitrary risk functions.

As described in the Introduction, a popular estimation strategy in this setting, especially if p is large, is to use an s/e (selection/estimation) procedure. With respect to (1.1), this consists of first using a variable selection procedure to estimate a subset of $\{\beta_2, \ldots, \beta_p\}$ by zeros (the intercept β_1 is always included), and second to estimate the remaining β_i's by least squares. An s/e procedure can be conveniently represented as follows. Let Γ be the set of all $1 \times p$ vectors of the form

$$\gamma = (\gamma_1, \ldots, \gamma_p), \tag{1.3}$$

where $\gamma_1 \equiv 1$ and $\gamma_i = 0$ or 1 for $i = 2, \ldots, p$. There are 2^{p-1} such vectors in Γ. An s/e procedure is then equivalent to first selecting $\gamma \in \Gamma$ such that $\gamma_i = 0$ if β_i is to be estimated by 0 and $\gamma_i = 1$ otherwise, and then estimating β by

$$\hat{\beta}_\gamma = (D_\gamma X'XD_\gamma)^{-1}D_\gamma X'Y, \tag{1.4a}$$

where

$$D_\gamma = \operatorname{diag}[\gamma] \tag{1.4b}$$

is the $p \times p$ diagonal matrix with diagonal elements γ. (A^{-1} is a generalized inverse of A.)

A variable selection procedure can then be represented as $\boldsymbol{\gamma}(Y, X) \equiv \boldsymbol{\gamma}$ (note bold $\boldsymbol{\gamma}$), a function from the data into Γ. Its consequent s/e procedure can then be represented by $\hat{\beta}_{\boldsymbol{\gamma}}$. Note that using least squares to estimate every component of β is the special case

$$\hat{\beta}_{\gamma_{\mathrm{LS}}} = (X'X)^{-1}X'Y, \qquad \gamma_{\mathrm{LS}} \equiv (1, 1, \ldots). \tag{1.5}$$

In this setting a natural criterion for evaluating an s/e procedure $\hat{\beta}_\gamma$ is the risk $R(\beta, \hat{\beta}_\gamma) = E_\beta|X\hat{\beta}_\gamma - X\beta|^2$. Unfortunately, the minimax point of view with this criterion may be unsatisfactory. Indeed, overall least squares $\hat{\beta}_{\gamma_{\mathrm{LS}}}$ is minimax for predictive risk. Thus, to be safe with respect to predictive risk, one should include all available predictors with γ_{LS} and use $\hat{\beta}_{\gamma_{\mathrm{LS}}}$. However, this perspective may encourage one to forego the large risk reductions available with s/e procedures. To remedy this deficiency, in the next section we propose the risk inflation criterion. It will be seen that the s/e procedure with minimum risk inflation is minimax with respect to a calibrated risk function. An alternative modification of the risk for related problems was considered by Cohen (1965).

2. The risk inflation criterion. The idea behind risk inflation is to calibrate the risk function to better reflect the potential gains from using an s/e procedure. Such gains are available in parameter space regions where many

of the β_i's are 0. In such regions we calibrate the risk function against the risk of the ideal (though unavailable) s/e procedure which correctly eliminates exactly the irrelevant X_i's. Of course, the best we can hope for with nonideal (but available) selection procedures is that such regions will be identified with high probability.

The risk inflation criterion is defined as follows. For each β, define

$$\eta(\beta) \equiv \eta = (1, \eta_2, \ldots, \eta_p) \in \Gamma, \tag{2.1}$$

where $\eta_i = I[\beta_i \neq 0]$. One can think of η as the "correct" value of γ. As mentioned above, the ideal (though unavailable) selection procedure would yield $\gamma \equiv \eta$, resulting in the ideal s/e procedure

$$\hat{\beta}_\eta = (D_\eta X'XD_\eta)^{-1} D_\eta X'Y, \tag{2.2}$$

which is just the "least squares estimator" based on exactly the "correct" predictors.

In practice, of course, η is unknown making $\hat{\beta}_\eta$ unavailable. A variable selection procedure γ may be thought of as an estimator of η, and the s/e procedure $\hat{\beta}_\gamma$ as a proxy for $\hat{\beta}_\eta$. For a particular β and its associated η [$=\eta(\beta)$], the increase in predictive risk from using $\hat{\beta}_\gamma$ instead of $\hat{\beta}_\eta$ would be $R(\beta, \hat{\beta}_\gamma)/R(\beta, \hat{\beta}_\eta)$. The risk inflation of γ is defined to be the maximum value of this ratio, namely

$$\mathrm{RI}(\gamma) \equiv \sup_\beta \{R(\beta, \hat{\beta}_\gamma)/R(\beta, \hat{\beta}_\eta)\}. \tag{2.3}$$

The risk inflation criterion (2.3) highlights the potential cost of using an s/e procedure. Small risk inflation corresponds to good performance with respect to $\hat{\beta}_\eta$. Note that the selection procedure with smallest risk inflation (if it exists) will be minimax with respect to the ratio function $R(\beta, \hat{\beta}_\gamma)/R(\beta, \hat{\beta}_\eta)$.

Up to this point, the definition of risk inflation in (2.3) is general and may be applied to any risk functions. In the special case of predictive risk (1.2), the risk inflation criterion is simplified by noting that the denominator is just the "least squares" risk

$$R(\beta, \hat{\beta}_\eta) \equiv |\eta|\sigma^2, \tag{2.4}$$

where $|\eta|$ is the number of nonzero components of η. Thus, risk inflation for predictive risk is

$$\mathrm{RI}(\gamma) = \sup_\beta R(\beta, \hat{\beta}_\gamma)/|\eta|\sigma^2. \tag{2.5}$$

A useful benchmark for comparison with other s/e procedures is the risk inflation of overall least squares $\hat{\beta}_{\gamma_{\mathrm{LS}}}$. Because it has constant predictive risk $R(\beta, \hat{\beta}_{\gamma_{\mathrm{LS}}}) = p\sigma^2$, its corresponding risk inflation is

$$\mathrm{RI}(\gamma_{\mathrm{LS}}) = \sup_\beta R(\beta, \hat{\beta}_{\gamma_{\mathrm{LS}}})/|\eta|\sigma^2 = \max_\eta p/|\eta| = p. \tag{2.6}$$

Of course, γ_{LS} would be safer than ignoring the data and using a fixed γ (i.e., arbitrarily excluding variables) since $RI(\gamma) = \infty$ unless $\gamma \equiv \gamma_{LS}$. However, as will be seen in subsequent sections, selection procedures based on the data can substantially improve on $RI(\gamma_{LS}) = p$.

A reasonable criticism of the risk inflation criterion (2.3) is that it calibrates against the risk of the ideal s/e procedure $\hat{\beta}_\eta$ which only excludes predictors whose coefficients are exactly 0. Since improved predictive performance can sometimes be obtained by excluding predictors with small but nonzero coefficients, a more realistic measure of risk inflation might be

$$\widetilde{RI}(\gamma) \equiv \sup_\beta \left\{ R(\beta, \hat{\beta}_\gamma) \Big/ \left[\inf_\gamma R(\beta, \hat{\beta}_\gamma) \right] \right\}. \tag{2.7}$$

With this criterion the risk of an s/e procedure is calibrated against the smallest risk achievable for β with an estimator of the form $\hat{\beta}_\gamma$ (with fixed γ), namely $\inf_\gamma R(\beta, \hat{\beta}_\gamma)$. As we show in Section 5, both RI and $\widetilde{RI}$ yield similar results.

Still other measures of risk inflation may be considered. For example, instead of a ratio measure, one might consider the minimax regret difference $\sup_\beta(R(\beta, \hat{\beta}_\gamma) - R(\beta, \hat{\beta}_\eta))$; see Venter and Steel (1992). This criterion may be unsatisfactory because for prediction risk it appears to allow the inclusion of a large fraction of the X's even when $\beta \equiv 0$.

3. A canonical variable selection procedure. In this section we describe a canonical form for a variable selection procedure which enables us to assess the risk inflation of the more commonly used selection procedures. We consider here the case of σ^2 known in order to better expose the main issues. It is seen below that the variable selection procedures AIC, C_p and BIC in this case are all of this form. In Section 6 it is shown that when σ^2 is unknown, the essential features of this canonical form and our analysis of it remain the same.

The general canonical selection procedure is defined as

$$\gamma_\Pi = \arg\min_{\gamma \in \Gamma} \left[SSE_\gamma + |\gamma| \sigma^2 \Pi \right], \tag{3.1a}$$

where $\Pi \geq 0$ is a prespecified constant,

$$SSE_\gamma \equiv |Y - X\hat{\beta}_\gamma|^2 \tag{3.1b}$$

and

$$|\gamma| \text{ is the number of nonzero components of } \gamma. \tag{3.1c}$$

Because $|\gamma|$ is the dimension of the model selected by γ, the procedure γ_Π selects that γ which minimizes the residual sum of squares SSE_γ penalized by $\sigma^2\Pi$ times the dimension of the model. We shall refer to Π as the dimensionality penalty of the procedure γ_Π.

Note that γ_Π depends in no essential way on the sample size n. This can be seen by writing

$$SSE_\gamma = \left| Y - X\hat{\beta}_{\gamma_{LS}} \right|^2 + \left| X\hat{\beta}_{\gamma_{LS}} - X\hat{\beta}_\gamma \right|^2,$$

which shows that only $|X\widehat{\beta}_{\gamma_{LS}} - X\widehat{\beta}_{\gamma}|^2$, which has $p - |\gamma|$ degrees of freedom, plays a role in determining γ_Π. We should also point out that in order to calculate γ_Π, it is necessary, in principle, to calculate SSE_γ for every γ. This is a characteristic of all the selection procedures considered in this paper. We plan to report elsewhere on alternative methods such as forward stepwise selection which have the computational advantage of examining fewer γ; see Miller (1990).

Intuition behind γ_Π is enhanced when $X'X$ is diagonal. In this case γ_Π may be expressed as

$$\gamma_\Pi = \{1, \gamma_2^*, \ldots, \gamma_p^*\} \quad \text{where } \gamma_i^* = I[\mathrm{SS}_i \geq \sigma^2\Pi] \tag{3.2a}$$

and

$$\mathrm{SS}_i = (X_i'Y)^2/(X_i'X_i), \qquad i = 2, \ldots, p. \tag{3.2b}$$

Here, γ_Π is the familiar stepwise selection, and Π is the F-to-enter or F-to-delete control parameter; see Miller (1990). Note also that here the computational burden for evaluating γ_Π is greatly reduced.

It is easy to see that for σ^2 known, the variable selection procedures AIC, C_p and BIC are all special cases of γ_Π in (3.1). We begin with the procedure proposed by Akaike (1974) which maximizes the criterion $\mathrm{AIC} = \log M_\gamma - |\gamma|$, where M_γ is the maximum likelihood of the model identified by γ. Akaike motivated this criterion as an estimate of the expected Kullback–Leibler information of the fitted model. For σ^2 known, Akaike's procedure is easily computed to be

$$\gamma_{\mathrm{AIC}} \equiv \arg\min_{\gamma} \mathrm{AIC}, \qquad \mathrm{AIC} = (1/2\sigma^2)\left[\mathrm{SSE}_\gamma + |\gamma|\sigma^2 2\right]. \tag{3.3}$$

Thus, γ_{AIC} is the special case of γ_Π with $\Pi = 2$.

The C_p procedure, attributed to Mallows (1973), is given by

$$\gamma_{C_p} \equiv \arg\min_{\gamma} C_p, \qquad C_p = \left[\mathrm{SSE}_\gamma/\sigma^2\right] - (n - 2|\gamma|) \tag{3.4a}$$

for σ^2 known. Mallows motivated C_p as an unbiased estimate for p, when all of the "correct" predictors had been selected. Interestingly, he recommended using graphical analysis of C_p plots for variable selection and cautioned against the pitfalls of using the automatic procedure γ_{C_p}. Nonetheless, it has become a popular practical criterion. Rewriting C_p as

$$C_p = \sigma^{-2}\left[\mathrm{SSE}_\gamma + |\gamma|\sigma^2 2\right] - n \tag{3.4b}$$

shows that γ_{C_p} is identical to γ_{AIC}, the special case of γ_Π with $\Pi = 2$. That AIC and C_p are the same in this context was noted by Stone (1977).

Schwarz (1978) proposed the selection procedure which maximizes the criterion $\mathrm{BIC} = \log M_\gamma - (1/2)|\gamma| \log n$, where M_γ is as in AIC. This criterion was motivated as a large-sample version of a Bayes procedure. For σ^2 known, Schwarz's procedure is

$$\gamma_{\mathrm{BIC}} \equiv \arg\min_{\gamma} \mathrm{BIC}, \qquad \mathrm{BIC} = (1/2\sigma^2)\left[\mathrm{SSE}_\gamma + |\gamma|\sigma^2(\log n)\right]. \tag{3.5}$$

Thus, γ_{BIC} is the special case of γ_Π with $\Pi = \log n$.

4. The risk inflation of γ_Π for $X'X$ diagonal. This section investigates properties of the risk inflation of γ_Π when $X'X$ is diagonal. Expressions are obtained which reveal the relationship between the risk inflation $\mathrm{RI}(\gamma_\Pi)$ and the dimensionality penalty Π. The risk inflation of the special cases AIC, C_p and BIC are then evaluated and compared. Finally, it is shown in Section 4.3 that $\Pi \approx 2\log p$ is optimal in the sense of minimizing the risk inflation of γ_Π. The risk inflation of γ_Π under general $X'X$ is investigated in Section 5.

4.1. *The predictive risk of $\hat{\beta}_{\gamma_\Pi}$.* In order to obtain the risk inflation of γ_Π, we require the risk of $\hat{\beta}_{\gamma_\Pi}$. We begin with the following expression for the predictive risk of a general s/e procedure. Expanding (1.2), it is straightforward to see that for $X'X$ diagonal

$$R(\beta, \hat{\beta}_\gamma) = V(\beta, \hat{\beta}_\gamma) + B(\beta, \hat{\beta}_\gamma), \tag{4.1a}$$

where

$$\begin{aligned} V(\beta, \hat{\beta}_\gamma) &= E_\beta \sum_{\gamma_i = 1} (X_i'\varepsilon)^2/|X_i|^2 = \sigma^2 + \sum_{i=2}^{p} E_\beta (X_i'\varepsilon)^2/|X_i|^2 I[\gamma_i = 1], \\ B(\beta, \hat{\beta}_\gamma) &= E_\beta \sum_{\gamma_i = 0} \left(|X_i|\beta_i\right)^2 = \sum_{i=2}^{p} \left(|X_i|\beta_i\right)^2 P[\gamma_i = 0]. \end{aligned} \tag{4.1b}$$

The terms in (4.1b) can be interpreted as follows. The "variance" component $V(\beta, \hat{\beta}_\gamma)$ accounts for the risk of estimating the selected components. (Note that the intercept is always selected.) The "bias" component $B(\beta, \hat{\beta}_\gamma)$ accounts for the nonzero components of β which were incorrectly set to 0.

As revealed by (3.2), the form of γ_Π also simplifies when $X'X$ is diagonal. To apply (4.1) to γ_Π in this case, note that SS_i in (3.2b) may be expressed as

$$\mathrm{SS}_i = \left(|X_i|\beta_i + (X_i'\varepsilon)/|X_i|\right)^2. \tag{4.2}$$

Because $(X_i'\varepsilon)/|X_i| = \sigma Z$ where $Z \sim N(0,1)$, we can write

$$\begin{aligned} E_\beta (X_i'\varepsilon)^2/|X_i|^2 I[\gamma_i^* = 1] &= \sigma^2 E\left[Z^2 I\left[\left(|X_i|\beta_i + \sigma Z\right)^2 > \sigma^2\Pi\right]\right], \\ \left(|X_i|\beta_i\right)^2 P[\gamma_i^* = 0] &= \left(|X_i|\beta_i\right)^2 P\left[\left(|X_i|\beta_i + \sigma Z\right)^2 \le \sigma^2\Pi\right]. \end{aligned} \tag{4.3}$$

Inserting (4.3) into (4.2) yields

$$R(\beta, \hat{\beta}_{\gamma_\Pi}) = \sigma^2 + \sigma^2 \sum_{i=2}^{p} R^*\left(|X_i|\beta_i/\sigma, \Pi\right), \tag{4.4a}$$

where

$$R^*(w, \Pi) \equiv E\left[Z^2 I\left[(w+Z)^2 > \Pi\right]\right] + w^2 P\left[(w+Z)^2 \le \Pi\right]. \tag{4.4b}$$

These expressions were previously considered by Mallows (1973) for $\Pi = 2$ and by Miller (1990).

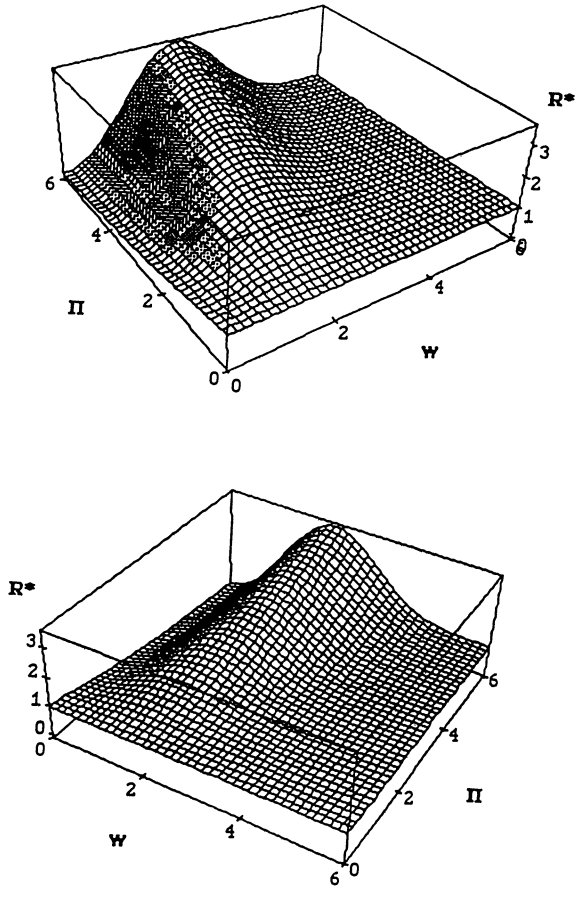

FIG. 1. *The risk surface.*

Note that each R^* term in (4.4) depends only on $w = |X_i|\beta_i/\sigma$ and Π. The first term in (4.4b) is the variance component and the second term is the bias component. Although not expressible in closed form, (4.4b) can be easily approximated numerically. Figure 1 shows the surface $R^*(w, \Pi)$. As will be proved later, it appears that the maximum occurs along a ridge, and is increasing as both w and Π increase. Slices of $R^*(w, \Pi)$ as a function of w for fixed $\Pi = 0, 1, 2, 3, 4$ are pictured in Figure 2. Note that when $\Pi = 0$, $R^* \equiv 1$ since in this case γ_Π is just overall least squares γ_{LS}. As Π increases, R^* decreases for small w but has an increasing maximum.

Finally, it should be noted that the only distributional assumption needed for (4.4b) and all of the results which follow is the normality of $(X_i'\varepsilon)/|X_i|$. By central limit theorem considerations, this assumption and hence virtually all of our results will be robust against many departures from the normality of ε.

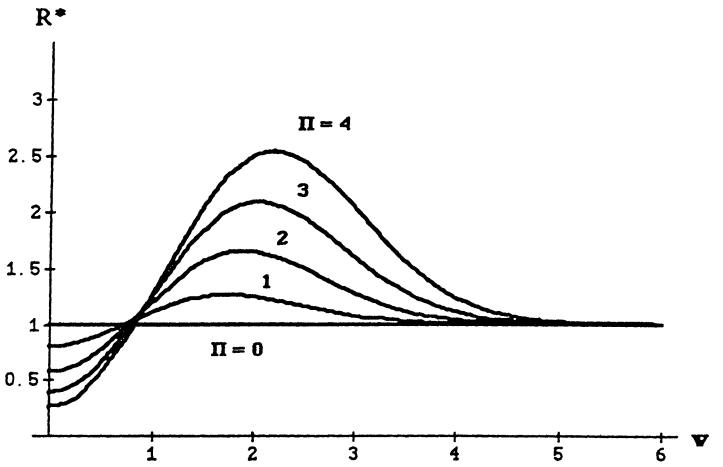

FIG. 2. $R^*(w, \Pi)$ *for various* Π.

4.2. *The risk inflation of* γ_Π. The calculation and subsequent analysis of the risk inflation $\mathrm{RI}(\gamma_\Pi)$ using (4.4) is facilitated by first calculating the partial risk inflation defined as

$$\mathrm{RI}(j, \gamma) \equiv \sup_{\beta \in B_j} R(\beta, \hat{\beta}_\gamma)/j\sigma^2 \quad \text{where } B_j = \{\beta: |\eta| = j\}, \tag{4.5}$$

the maximum risk over B_j, the set of β's with exactly j nonzero components. From (4.4), it follows that

$$\mathrm{RI}(j, \gamma_\Pi) = (1/j)\left[1 + (p - j)R^*(0, \Pi) + (j - 1) \sup_w R^*(w, \Pi)\right], \tag{4.6}$$

a function only of Π and j. The risk inflation of γ_Π is now easily obtained as

$$\begin{aligned} \mathrm{RI}(\gamma_\Pi) &= \max_j \mathrm{RI}(j, \gamma_\Pi) = \mathrm{RI}(1, \gamma_\Pi) \vee \mathrm{RI}(p, \gamma_\Pi) \\ &= \left[1 + (p - 1)R^*(0, \Pi)\right] \vee \left[1/p + (1 - 1/p) \sup_w R^*(w, \Pi)\right], \end{aligned} \tag{4.7}$$

where $\vee$ is the maximum operator. The two crucial quantities in $\mathrm{RI}(j, \gamma_\Pi)$ and $\mathrm{RI}(\gamma_\Pi)$ are $R^*(0, \Pi)$ and $\sup_w R^*(w, \Pi)$.

The quantity $R^*(0, \Pi)$ accounts for the error estimating $\beta_i = 0$ when $\hat{\beta}_{\gamma_\Pi}$ is used. From (4.4b), one can see that $R^*(0, \Pi)$ is composed exclusively of estimation risk. For computational purposes, $R^*(0, \Pi)$ can be computed directly by

$$R^*(0, \Pi) = 2\left[\sqrt{\Pi}\phi(\sqrt{\Pi}) + \Phi(-\sqrt{\Pi})\right], \tag{4.8}$$

where ϕ and Φ are the standard normal pdf and cdf. Figure 3 displays $R^*(0, \Pi)$ which decreases exponentially from $R^*(0, 0) = 1$ as Π increases. This decrease

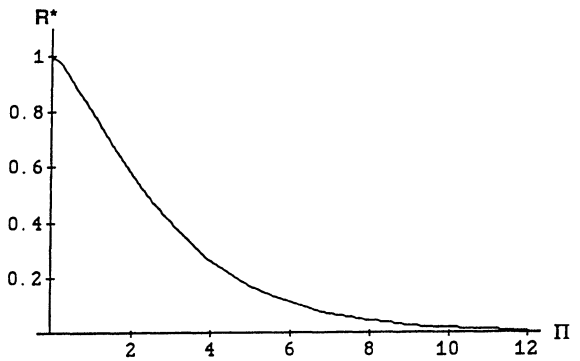

FIG. 3. *The component risk at* $\beta_i = 0$: $R^*(0, \Pi)$.

is a manifestation of the feature that $\beta_i = 0$ is less likely to be "incorrectly" estimated when Π is large. Some numerical values of $R^*(0, \Pi)$ are presented in Table 1.

The quantity $\sup_w R^*(w, \Pi)$ accounts for that w, when $\hat{\beta}_{\gamma_\Pi}$ is used, which yields the worst possible combination of estimation risk and bias risk in (4.4b). Although a closed form expression for $\sup_w R^*(w, \Pi)$ is unavailable, it follows from Lemma A.1 in the Appendix that

$$\Pi - o(\Pi) < \sup_w R^*(w, \Pi) < \Pi + 1. \tag{4.9}$$

(Lemma A.1 actually obtains an even sharper left-hand bound.) For practical purposes, $\sup_w R^*(w, \Pi)$ can be easily obtained numerically. Figure 4 displays this maximum risk, which increases almost linearly from $\sup_w R^*(w, 0) = 1$ as Π increases. This increase is a manifestation of the feature that bias risk increases faster than estimation risk as Π increases. Some numerical values of $\sup_w R^*(w, \Pi)$ are presented in Table 1.

Based on the values in Table 1, it is possible to calculate the $\mathrm{RI}(\gamma_\Pi)$ using (4.7) for a variety of Π. For p and Π large, it is perhaps more revealing to consider the approximation of (4.7),

$$\mathrm{RI}(\gamma_\Pi) \approx p2\sqrt{\Pi}\phi(\sqrt{\Pi}) \vee \Pi, \tag{4.10a}$$

which is based on

$$\mathrm{RI}(1, \gamma_\Pi) \approx pR^*(0, \Pi) \approx p2\sqrt{\Pi}\phi(\sqrt{\Pi}) \tag{4.10b}$$

and

$$\mathrm{RI}(p, \gamma_\Pi) \approx \sup_w R^*(w, \Pi) \approx \Pi, \tag{4.10c}$$

which in turn makes use of (4.6), (4.8) and (4.9). The approximate values of $\mathrm{RI}(\gamma_\Pi)$ using (4.10) are displayed in Table 2 for $\Pi = 1, 2, \log n$. As will be seen in Section 6, the choice $\Pi = 1$ corresponds to maximizing adjusted R^2. The choices $\Pi = 2$ and $\Pi = \log n$ correspond to AIC/C_p and BIC, respectively, as pointed out

TABLE 1
The risk components

Π	$R^*(0, \Pi)$	$\sup_w R^*(w, \Pi)$
0	1	1
.5	.918	1.099
1.0	.801	1.260
1.5	.682	1.448
2.0	.572	1.650
2.5	.475	1.862
3.0	.391	2.082
3.5	.320	2.307
4.0	.261	2.537
4.5	.212	2.772
5.0	.171	3.011
6.0	.111	3.500
7.0	.0718	4.000
8.0	.0460	4.517
9.0	.0292	5.045
10.	.01856	5.581
15.	.001816	8.399
20.	.000169	11.39
25.	.0000154	14.52
30.	.00000138	17.76
Π Large	$\approx 2\sqrt{\Pi}\Phi(\Pi)$	$\approx \Pi$

in Section 3. Note that all of these offer substantially less risk inflation than overall γ_{LS}, corresponding to $\Pi = 0$, and which has $RI(\gamma_{LS}) = p$. Note also that for $\Pi = 1, 2, \log n$, and large p, $RI(\gamma_\Pi)$ is decreasing in Π.

4.3. *For X'X diagonal, $\Pi \approx 2 \log p$ is optimal.* Although AIC, C_p, BIC and a variety of other γ_Π's may offer smaller risk inflation than γ_{LS}, the issue arises as to which Π minimizes $RI(\gamma_\Pi)$. Consider Figure 5 which displays the partial risk inflations $RI(1, \gamma_\Pi)$ and $RI(p, \gamma_\Pi)$ as functions of Π for $p = 2$, 10, 100, 200. From (4.7), $RI(\gamma_\Pi)$ is the maximum of these two curves at each Π. For smaller Π, $RI(\gamma_\Pi) = RI(1, \gamma_\Pi)$ which appears to be exponentially decreasing, and for larger Π, $RI(\gamma_\Pi) = RI(p, \gamma_\Pi)$, which appears to be linearly increasing. The minimum occurs at the intersection of these two curves.

Using the approximation (4.10), the Π which yields the minimum RI approximately satisfies $p2\sqrt{\Pi}\phi(\sqrt{\Pi}) = \Pi$. Although more precise solutions to this equality can be obtained, it appears that

$$\Pi = 2 \log p \tag{4.11}$$

is a reasonable candidate, and more than adequate for practical purposes. Indeed, Table 3 provides strong support for the use of $\Pi = 2 \log p$ as an approximation to the optimal Π. The first two columns of Table 3 show remarkable agreement for a wide range of values of p, between $\Pi = 2 \log p$ and the optimal Π

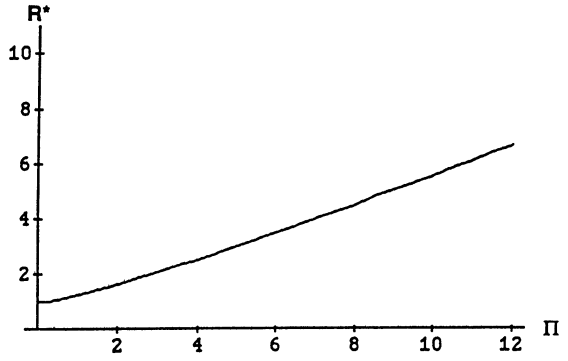

FIG. 4. *The worst possible component risk*: $\sup R^*(w, \Pi)$.

(which was obtained numerically). The last two columns also show remarkable agreement between the risk inflation value corresponding to $\Pi = 2\log p$ and the smallest available risk inflation available with a procedure of the form γ_Π.

Because the optimal Π satisfies $\mathrm{RI}(\gamma_\Pi) = \mathrm{RI}(1, \gamma_\Pi) = \mathrm{RI}(p, \gamma_\Pi)$, the approximation (4.10) suggests that for optimal Π, $\mathrm{RI}(\gamma_\Pi) \approx \sup_w R^*(w, \Pi) \approx \Pi$. Thus, we might expect that for p large

$$\mathrm{RI}(\gamma_{2\log p}) \approx 2\log p. \tag{4.12}$$

The following result shows that for $X'X$ diagonal, this approximation improves as $p \to \infty$, and that $2\log p$ is the smallest possible risk inflation for *any* selection procedure γ, not just those of the form γ_Π. The powerful implication of this result is that for $X'X$ diagonal, $\gamma_{2\log p}$ is asymptotically optimal (as $p \to \infty$) with respect to risk inflation within the class of all variable selection procedures. The optimal bound $2\log p$ has also recently been obtained in a related wavelet model selection problem by Donoho and Johnstone (1994).

TABLE 2
The risk inflation of various procedures (The $X'X$ diagonal case)

Method	Π	$R^*(0, \Pi)$	$\sup_w R^*(w, \Pi)$	Risk inflation
LS	0	1	1	p
max adj R^2	1	.801	1.26	$\approx p(.801)$
AIC/C_p	2	.573	1.65	$\approx p(.573)$
BIC	$\log n$	$\approx\sqrt{(2\log n)/(\pi n)}$	$\approx \log n$	$\approx \log n$ if $p \ll \sqrt{n}$ $\approx\sqrt{(2\log n)/(\pi n)}$ if $p \gg \sqrt{n}$
General γ_Π	Π	$\approx 2\sqrt{\Pi}\phi(\Pi)$	$\approx\Pi$	$\approx p 2\sqrt{\Pi}\phi(\Pi) \vee \Pi$
$\gamma_{2\log p}$	$2\log p$	$\approx\sqrt{(4\log p)/(\pi p)^2}$	$\approx 2\log p$	$\approx 2\log p$
General γ				$\geq 2\log p - o(\log p)$

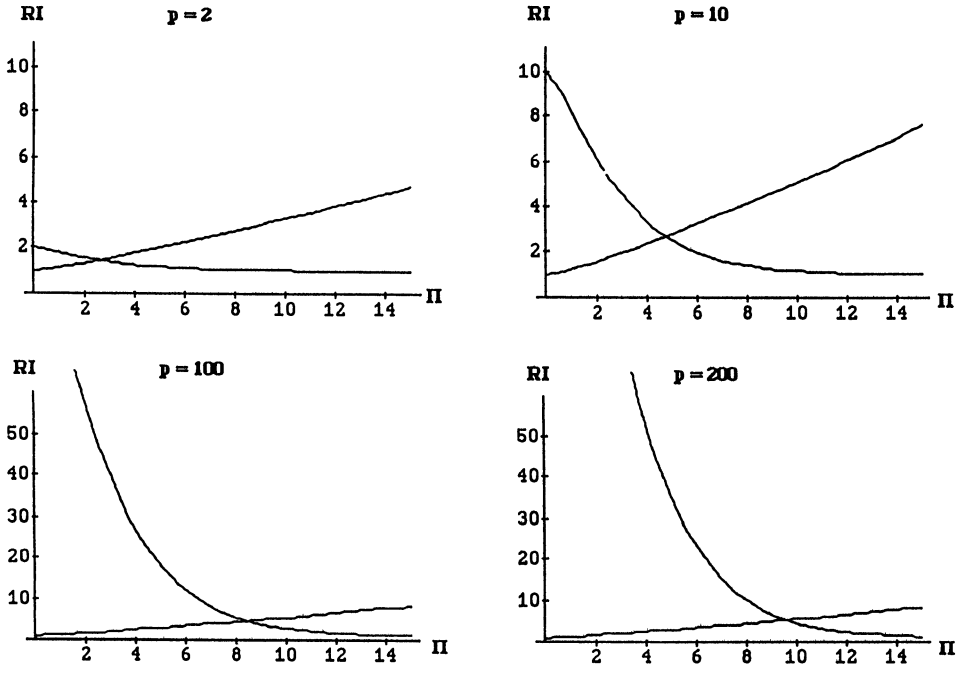

FIG. 5. $RI(\gamma_\Pi) = RI(1, \gamma_\Pi) \wedge RI(p, \gamma_\Pi)$. *Note: In each figure* $RI(1, \gamma_\Pi)$ *is the downward sloping curve and* $RI(p, \gamma_\Pi)$ *is the upward sloping curve.*

THEOREM 4.1. *For X'X diagonal,*

(i) $$\mathrm{RI}(\gamma_{2\log p}) < 1 + 2\log p.$$

(ii) *For any* γ, $$\mathrm{RI}(\gamma) \geq 2\log p - o(\log p).$$

PROOF. To prove (i), from (4.7) it suffices to show that $\mathrm{RI}(j, \gamma_{2\log p}) < 1 + 2\log p$ for $j = 1$ and $j = p$. By (4.6) and (4.8),

$$\mathrm{RI}(1, \gamma_{2\log p}) < 1 + pR^*(0, 2\log p) < +p2\sqrt{2\log p}\,\phi(\sqrt{2\log p}) < 1 + \sqrt{(4\log p)/\pi} < 1 + 2\log p.$$

By (4.6) and (4.9)

$$\mathrm{RI}(p, \gamma_{2\log p}) < \sup_w R^*(w, 2\log p) < 1 + 2\log p.$$

TABLE 3
The risk inflation when $\Pi = 2 \log p$ *(The X'X diagnoal case)*

p	$\Pi = 2 \log p$	BEST Π	RI($\gamma_{2 \log p}$)	inf RI
1	0	—	1.00	1.00
2	1.39	2.62	1.71	1.46
3	2.20	3.11	2.06	1.75
4	2.77	3.48	2.28	1.97
5	3.22	3.78	2.44	2.15
6	3.58	4.03	2.55	2.29
7	3.89	4.25	2.64	2.42
8	4.16	4.44	2.71	2.53
9	4.39	4.61	2.77	2.62
10	4.61	4.77	2.82	2.71
15	5.42	5.38	3.07	3.05
20	5.99	5.83	3.37	3.29
30	6.80	6.47	3.81	3.64
50	7.82	7.30	4.36	4.09
100	9.21	8.46	5.12	4.72
200	10.60	9.64	5.88	5.37
500	12.43	11.24	6.91	6.25
1,000	13.82	12.47	7.71	6.94
10,000	18.42	16.65	10.43	9.37
100,000	23.03	20.92	13.28	11.96
1,000,000	27.63	25.25	16.22	14.69
10,000,000	32.24	29.63	19.24	17.52

To prove (ii),

$$\begin{aligned}
\mathrm{RI}(\gamma) &= \max_j \mathrm{RI}(j, \gamma) \\
&= \max_j \left(j\sigma^2\right)^{-1} \sup_{\beta \in B_j} R(\beta, \hat{\beta}_\gamma) \\
&\geq \max_j \left(j\sigma^2\right)^{-1} \sigma^2 \left[2(j-1)\log p - o(\log p)\right] \\
&\geq \max_j \left((j-1)/j\right)\left[2\log p - o(\log p)\right] \geq 2\log p - o(\log p),
\end{aligned}$$

where the first inequality makes use of $\sup_{\beta \in B_j} R(\beta, \hat{\beta}_\gamma) \geq \sigma^2[2(j-1)\log p - o(\log p)]$, a special case of Lemma A.2 in the Appendix. □

A feature not brought out in the proof of Theorem 4.1 is that Lemma A.2 shows

$$\sup_{\beta \in B_j} R(\beta, \hat{\beta}) \geq \sigma^2 \left[2(j-1)\log p - o(\log p)\right]. \tag{4.13}$$

for *any* estimator $\hat{\beta}$. As a result, the lower bound provided by (ii) above applies to the predictive risk of any estimator $\hat{\beta}$, not just an s/e estimator. Thus, going outside the class of s/e estimators offers no improvement in risk inflation.

As discussed in Section 2, a reasonable alternative definition of risk inflation is $\widetilde{\mathrm{RI}}$ in (2.7). The following result for $\widetilde{\mathrm{RI}}$ is very similar to Theorem 4.1 and shows that $\Pi \approx 2\log p$ is also optimal for $\widetilde{\mathrm{RI}}$.

THEOREM 4.2. *For $X'X$ diagonal,*

(i) $$\widetilde{\mathrm{RI}}(\gamma_\Pi) < 2\log p + o(\log p) \quad \text{for } \Pi = 2\log p + 2\sqrt{2\log p} + 1.$$

(ii) $$\text{For any } \gamma,\ \widetilde{\mathrm{RI}}(\gamma) \geq 2\log p - o(\log p).$$

PROOF. It follows from (2.7) and (4.4a) that

$$\widetilde{\mathrm{RI}}(\gamma_\Pi) = \sup_{w_2,\ldots,w_p} \frac{1 + \sum_{i=2}^{p} R^*(w_1, \Pi)}{1 + \sum_{i=2}^{p} \min(w_i^2, 1)}. \tag{4.14}$$

It can be shown using (4.4b), (4.8) and (4.9) that for $\Pi = 2\log p + 2\sqrt{2\log p} + 1$,

$$\begin{aligned} R^*(w,\Pi) &= w^2 + o(1/p) && \text{for } w^2 \leq 1, \\ R^*(w,\Pi) &\leq \Pi + 1 && \text{for } w^2 > 1; \end{aligned} \tag{4.15}$$

part (i) follows by combining (4.14) and (4.15). Part (ii) follows immediately from (ii) of Theorem 4.1 and the obvious fact that for any γ, $\widetilde{\mathrm{RI}}(\gamma) \geq \mathrm{RI}(\gamma)$. □

5. Risk inflation bounds for general $X'X$. In the general case where $X'X$ is not necessarily diagonal, simple expressions for $\mathrm{RI}(\gamma_\Pi)$ seem unavailable. Furthermore, the optimal value of Π for γ_Π depends on the correlation structure of $X'X$. The following result shows however, that useful upper bounds for the risk can be obtained, and that for large Π, these bounds are similar to the diagonal case.

THEOREM 5.1. *Define $\xi = \sqrt{\Pi}e^{(1-\Pi)/2}$. For any γ_Π,*

(i) $$R(\beta, \hat{\beta}_{\gamma_\Pi}) \leq p\sigma^2(\sqrt{\Pi} + 1)^2;$$

(ii) $$R(\beta, \hat{\beta}_{\gamma_\Pi}) \leq 2\sigma^2|\eta|(\Pi + 1) + \sigma^2 4\sqrt{2}\left(\Pi/(\Pi - 1)\right)^2 e^{p\xi}\sqrt{p\xi}\,.$$

PROOF. See the Appendix.

The risk bound in (i) above is useful for Π small (about $\Pi < 2\log p - 2\log\log p$), whereas the bound in (ii) is useful for Π large (about $\Pi > 2\log p + 2\log\log p$). The second bound is particularly useful for large p as indicated by the following corollary.

COROLLARY 5.2. *If* $\Pi > 2\log p + 2\log\log p$, *then as* $p \to \infty$,

(i) $$R(\beta, \hat{\beta}_{\gamma_\Pi}) \le \sigma^2 2|\eta|(\Pi + 1) + o(1).$$

(ii) $$\mathrm{RI}(\gamma_\Pi) \le 2(\Pi + 1) + o(1).$$

PROOF. Part (i) follows from (ii) of Theorem 5.1 and the observation that $\xi = \sqrt{\Pi}\, e^{(1-\Pi)/2} = o(p^{-1})$ as long as $\Pi \ge 2\log p + (1+\varepsilon)\log\log p$ for some $\varepsilon > 0$; (ii) follows immediately from (i). □

For general $X'X$, the smallest bound on $\mathrm{RI}(\gamma_\Pi)$ obtainable from Corollary 5.2 occurs for $\Pi \approx 2\log p$, ($2\log\log p$ will be small), which for large p is about $4\log p$, twice that in the diagonal case. The next result shows that in fact this choice is asymptotically best for at least the worst possible X. In this sense, $\gamma_{2\log p}$ is asymptotically safest. Of course, for particular known X, a better choice of Π may be obtained numerically.

THEOREM 5.3. *For any* Π, $\sup_X \mathrm{RI}(\gamma_\Pi) \ge 4\log p - o(\log p)$.

PROOF. It follows from Lemma A.5 in the Appendix that X and β can be chosen such that for any integer $1 \le j \le p$, $\sup_X \mathrm{RI}(\gamma_\Pi) \ge ((j-1)/j)4\log p - o(\log p))$. □

6. Unknown σ^2. In this section we consider the case of unknown σ^2. In this case, many of the popular selection procedures including AIC, C_p and BIC are of the canonical form

$$\gamma_{\hat{\Pi}} = \arg\min_{\gamma \in \Gamma} \left[\mathrm{SSE}_\gamma + |\gamma|\sigma^2\hat{\Pi}\right], \tag{6.1}$$

where $\hat{\Pi} \ge 0$ is a stochastic dimensionality penalty and SSE_γ and $|\gamma|$ are as in (3.1). As opposed to Π in the canonical procedure γ_Π in (3.1), $\hat{\Pi}$ is not a prespecified constant but rather a random variable which depends on the data. However, for the procedures we examine, $\hat{\Pi}$ will converge to some Π in some sense as $n \to \infty$, so that for large n, $\gamma_{\hat{\Pi}}$ will behave like the corresponding γ_Π.

For example, when σ^2 is unknown, the C_p procedure is modified by substituting an estimator of σ^2. This estimator is typically

$$\hat{\sigma}^2_{\mathrm{LS}} \equiv |Y - X\hat{\beta}_{\gamma_{\mathrm{LS}}}|/(n-p), \tag{6.2a}$$

yielding

$$\gamma_{C_p} \equiv \arg\min_{\gamma} C_p, \qquad C_p = \left[\mathrm{SSE}_\gamma/\hat{\sigma}^2_{\mathrm{LS}}\right] - \left(n - 2|\gamma|\right). \tag{6.2b}$$

Rewriting

$$C_p = \hat{\sigma}^{-2}_{\mathrm{LS}}\left[\mathrm{SSE}_\gamma + |\gamma|\hat{\sigma}^2_{\mathrm{LS}}2\right] - n \tag{6.2c}$$

shows that using γ_{C_p} corresponds to $\gamma_{\hat{\Pi}}$ with $\hat{\Pi} = 2\,\hat{\sigma}^{\,2}_{\mathrm{LS}}/\sigma^2$. Here $\hat{\Pi} \to 2$, so that for large n, this procedure agrees with γ_{C_p} in (3.4a) for σ^2 known.

Similarly, the procedures AIC and BIC are shown to be of the form (6.1). When σ^2 is unknown, Akaike's procedure is

$$\gamma_{\mathrm{AIC}} \equiv \arg\min_{\gamma} \mathrm{AIC}, \qquad \mathrm{AIC} = (n/2)\log\left(\mathrm{SSE}_\gamma/n\right) + |\gamma|. \tag{6.3}$$

Reexpression of (6.3) shows that γ_{AIC} is $\gamma_{\hat{\Pi}}$ with $\hat{\Pi} \equiv 2 + O[(\mathrm{SSE}_{\gamma_{\mathrm{AIC}}}/n\sigma^2) - 1]^2 \to 2$, agreeing with γ_{AIC} in (3.3) for large n. When σ^2 is unknown, the BIC procedure is

$$\gamma_{\mathrm{BIC}} \equiv \arg\min_{\gamma} \mathrm{BIC}, \qquad \mathrm{BIC} = (n/2)\log\left(\mathrm{SSE}_\gamma/n\right) + (1/2)|\gamma|(\log n). \tag{6.4}$$

Reexpression of (6.4) shows that γ_{BIC} is $\gamma_{\hat{\Pi}}$ with

$$\hat{\Pi} = \log n + O[(\mathrm{SSE}_{\gamma_{\mathrm{BIC}}}/n\sigma^2) - 1]^2 \to \log n,$$

agreeing with γ_{BIC} in (3.5) for large n.

It is useful to note that other commonly used selection procedures can only be expressed in the form (6.1) and not (3.1). For example, consider the well-known procedure of maximizing adjusted R^2 [Theil (1961)] or equivalently minimizing the residual variance estimate, namely

$$\gamma_{aR^2} \equiv \arg\max_{\gamma}\left[1 - \frac{\mathrm{SSE}_\gamma/(n - |\gamma|)}{\mathrm{SS}_{\mathrm{TOT}}/(n-1)}\right] \equiv \arg\min_{\gamma} \mathrm{SSE}_\gamma/(n - |\gamma|). \tag{6.5a}$$

Rewriting

$$\mathrm{SSE}_\gamma/(n - |\gamma|) = n^{-1}\left[\mathrm{SSE}_\gamma + |\gamma|\left(\mathrm{SSE}_\gamma/(n - |\gamma|)\right)\right] \tag{6.5b}$$

shows that γ_{aR^2} corresponds to using $\gamma_{\hat{\Pi}}$ with $\hat{\Pi} = \mathrm{SSE}_{\gamma_{aR^2}}/(n - |\gamma_{aR^2}|)\sigma^2$. Here $\hat{\Pi} \to 1$, so that for large n, maximizing adjusted R^2 corresponds to using γ_Π with $\Pi = 1$.

All of these procedures have stochastic dimensionality penalties which are relatively stable and converge (in some sense) as $n \to \infty$. Similarly, many other popular variable selection procedures, such as those mentioned the Introduction, can also be put in the form $\gamma_{\hat{\Pi}}$, with similarly stable and convergent $\hat{\Pi}$.

Effectively, the main difference between all of these procedures has to do with the relative sizes of the limits of their respective $\widehat{\Pi}$'s.

The following result shows that all of our previous results about the risk and risk inflation of procedures of the form γ_{Π} are limiting cases of similar results for the corresponding $\gamma_{\widehat{\Pi}}$. More precisely, as long as $\widehat{\Pi}$ is independent of $\widehat{\beta}_{\gamma_{\mathrm{LS}}}$ for all γ and $\widehat{\Pi}$ converges in L_1 to some Π_0, then $\gamma_{\widehat{\Pi}}$ is asymptotically equivalent to γ_{Π_0} in risk and risk inflation.

THEOREM 6.1. *Suppose that $\widehat{\Pi}_k$ is independent of $\widehat{\beta}_{\gamma_{\mathrm{LS}}}$ and that for some $\Pi_0, E|\widehat{\Pi}_k - \Pi_0| \to 0$ as $k \to \infty$. Then*

(i) $$\sup_{\beta} \left|R\left(\beta, \widehat{\beta}_{\gamma_{\widehat{\Pi}_k}}\right) - R\left(\beta, \widehat{\beta}_{\gamma_{\Pi_0}}\right)\right| \to 0.$$

(ii) $$\mathrm{RI}\left(\gamma_{\widehat{\Pi}_k}\right) \to \mathrm{RI}\left(\gamma_{\Pi_0}\right).$$

PROOF. See the Appendix.

7. Conclusion: a new variable selection procedure. Because of Theorem 6.1, when σ^2 is unknown and $(n - p)$ is reasonably large, we recommend the selection procedure

$$\gamma_{\mathrm{RIC}} = \arg\min_{\gamma \in \Gamma} \left[\mathrm{SSE}_{\gamma} + |\gamma| \widehat{\sigma}^2_{\mathrm{LS}} (2 \log p)\right], \tag{7.1}$$

where we have labeled RIC for risk inflation criterion. This procedure is the special case of $\gamma_{\widehat{\Pi}}$ in (6.1) with $\widehat{\Pi} = \widehat{\sigma}^2_{\mathrm{LS}} (2 \log p)/\sigma^2$. As is well known, the unbiased variance estimate $\widehat{\sigma}^2_{\mathrm{LS}}$ in (6.2a) is independent of the overall least squares estimate $\widehat{\beta}_{\gamma_{\mathrm{LS}}}$. Furthermore, $E|\widehat{\Pi} - 2 \log p| \to 0$ as $n \to \infty$. Thus, application of Theorem 6.1 to γ_{RIC} shows that as n increases, the risk properties of γ_{RIC} converge to those of γ_{Π_0} with $\Pi_0 = 2 \log p$.

Appealing to the results of Section 4, the risk inflation of γ_{RIC} will be asymptotically close to optimal in the orthogonal case and distributionally robust.

APPENDIX

To improve the readability and flow of the main text, the proofs of the main theorems, as well as some necessary lemmas, have been placed in this appendix. However, as elaborated in the Introduction, many of these results may also be of independent interest and useful elsewhere.

We begin with the following result which yields the bounds on $\sup_w R^*(w, \Pi)$ stated in (4.9) and is key to the proof of (i) of Theorem 4.1.

LEMMA A.1. $\Pi - 2\sqrt{2\Pi \log \Pi} + o(\sqrt{2\Pi \log \Pi}) < \sup_w R^*(w, \Pi) < 1 + \Pi$.

PROOF. Let $Z \sim N(0,1)$ throughout. Let us first show the right-hand inequality. We will show this for $\Pi \geq 12$. It is straightforward to strengthen this proof for smaller Π, although for small Π, the inequality is obvious from Figure 4. Without loss of generality, assume $w \geq 0$. If $w^2 < \Pi$,

$$R^*(w, \Pi) \leq 1 + w^2 P\big[(w+Z)^2 \leq \Pi\big] \leq 1 + w^2 < 1 + \Pi.$$

If $w^2 \geq \Pi$,

$$\begin{aligned} R^*(w, \Pi) &\leq 1 + w^2 P\big[(w+Z)^2 \leq \Pi\big] \\ &\leq 1 + w^2 \Phi(\sqrt{\Pi} - w) \\ &= 1 + \big(x^2 + 2x\sqrt{\Pi} + \Pi\big)\Phi(-x) \qquad \text{(A.1a)} \\ &\leq 1 + \Pi/2 + \big(x^2 + 2x\sqrt{\Pi}\big)\Phi(-x) \qquad \text{(A.1b)} \\ &< 1 + \Pi/2 + (x + 2\sqrt{\Pi})\phi(x) \qquad \text{(A.1c)} \\ &< 1 + \Pi/2 + 1 + \Pi/4 \leq 1 + \Pi, \qquad \text{(A.1d)} \end{aligned}$$

where (A.1a) makes use of $x = w - \sqrt{\Pi}$, (A.1b) makes use of $\Phi(-x) \leq 1/2$ for $x \geq 0$, (A.1c) makes use of $\Phi(-x) < \phi(x)/x$ and (A.1d) makes use of $x\phi(x) < 1$ and $2\sqrt{\Pi}\phi(x) < \Pi/4$ if $\Pi \geq 12$. We now proceed to show the left-hand inequality:

$$\begin{aligned} \sup_w R^*(w, \Pi) &\geq \sup_w w^2 P(Z \leq \sqrt{\Pi} - w) \qquad \text{(A.2a)} \\ &\geq \sup_{x \geq 0} \big(\Pi - 2x\sqrt{\Pi} + x^2\big)\Phi(x) \\ &\geq (\Pi - 2\sqrt{2\Pi \log \Pi} + 2\log \Pi)\Phi(\sqrt{2\log \Pi}) \\ &> \Pi\big(1 - 2\sqrt{2\log \Pi/\Pi} + 2\log \Pi/\Pi\big)\big(1 - 1/\Pi\big) \qquad \text{(A.2b)} \\ &= \Pi - 2\sqrt{2\Pi \log \Pi} + o(\sqrt{2\Pi \log \Pi}), \end{aligned}$$

where (A.2a) uses $E[Z^2 I[(w+Z)^2 \geq \Pi]] \geq w^2 P[Z \leq -w - \sqrt{\Pi}] + E[Z^2 I\ [Z > w - \sqrt{\Pi}]] \geq w^2 P[Z \leq -w - \sqrt{\Pi}]$, and (A.2b) follows from $\Phi(\sqrt{2\log \Pi}) > 1 - \phi(\sqrt{2\log \Pi})/\sqrt{2\log \Pi} > 1 - 1/\Pi$. □

The next result is the cornerstone of the proof of (ii) of Theorem 4.1.

LEMMA A.2. *For any estimator $\hat{\beta}$, when $X'X$ is diagonal,*

$$\sup_{\beta \in B_j} R(\beta, \hat{\beta}) \geq \sigma^2\big[2(j-1)\log p - o(\log p)\big].$$

PROOF. We assume wlog that $X'X = I$ and $\sigma^2 = 1$. Furthermore, by suitable rotation we can assume that $Y_i \sim N(\beta_i, 1)$, $i = 1, \ldots, p$, and $Y_i \sim N(0, 1)$, $i = p + 1, \ldots, n$, all independently. Thus, we can assume that $\hat{\beta}$ depends only on $Y_1, \ldots, Y_p$. Finally, for notational convenience we let $k = j - 1$ throughout. Define $A_{k,a} \subset B_j$ by

$$A_{k,a} = \{\beta: \beta_1 = 0, \beta_i = 0 \text{ or } a \text{ for } i = 2, \ldots, p, \text{ and } (\#\beta_i \neq 0) = k\}.$$

Then for any a,

$$\sup_{\beta \in B_j} R(\beta, \hat{\beta}) \geq \max_{\beta \in A_{k,a}} R(\beta, \hat{\beta}). \tag{A.3}$$

We now proceed to show that there exists a sequence $\{a_p\}$ (to be specified below), such that

$$\max_{\beta \in A_{k,a_p}} R(\beta, \hat{\beta}) \geq 2k \log p - o(\log p). \tag{A.4}$$

Coupled with (A.3), this will establish the desired result.

From here on, suppose $\beta \in A_{k,a_p}$. Since $\beta_1 = 0$ for such β, we set $\hat{\beta}_1 \equiv 0$. Our construction of $\{a_p\}$ to satisfy (A.4) now proceeds by augmenting the data $Y_2, \ldots, Y_p$ with a new set of random variables. This is done in such a way that we are then able to extract from the augmented data a set of sufficient statistics for the β_i's. It is then shown that the optimal estimator based on these sufficient statistics (which dominates any estimator based only on $Y_2, \ldots, Y_p$) has risk bounded below by the r.h.s of (A.4).

We begin by constructing 0, 1 random variables $I_2, \ldots, I_p$ (depending on $\beta_2, \ldots, \beta_p$) such that when $I_i = 1$, Y_i carries no information about β_i. Consider another sequence $\{\alpha_p\}$ (to be specified below). When $\beta_i = a_p$, let

$$I_i = I[Y_i \leq \alpha_p].$$

For such i, $Y_i \mid (I_i = 1)$ is $N(a_p, 1)$ truncated at α_p. When $\beta_i = 0$, let

$$P(I_i = 1 \mid Y_i) = I[Y_i \leq \alpha_p] \frac{\phi(\alpha_p)}{\phi(\alpha_p - a_p)} \frac{\phi(Y_i - a_p)}{\phi(Y_i)}.$$

Using a rejection sampling argument, it follows that here too, $Y_i \mid (I_j = 1)$ is $N(a_p, 1)$ truncated at α_p. Thus, when $I_i = 1$ the distribution of Y_i does not depend on β_i.

Now define the p indicator variables

$$J = \begin{cases} 1, & \text{if for some } i, Y_i \geq \alpha_p \text{ and } \beta_i = a_p, \\ 0, & \text{otherwise,} \end{cases}$$

and

$$I_i' = \begin{cases} I_i, & \text{if } J = 0, \\ I[\beta_i \neq 0], & \text{if } J = 1. \end{cases}$$

Note that $I_i' = 0$ implies that $\beta_i = 0$. When $I_i' = 1$ and $J = 1$, then $\beta_i = a_p$. However, when $J = 0$, the conditional distribution of $Y_i \mid I_i'$ does not depend on β_i. Thus (I_i', J) is sufficient for β_i.

As a consequence of this sufficiency, any estimator $\hat{\beta}(Y_2, \ldots, Y_p)$ which is a function of $Y_2, \ldots, Y_p$ can be dominated in risk by an estimator $\hat{\beta}'(I_2', \ldots, I_p', J)$ which is a function only of $I_2', \ldots, I_p', J$. In particular, it is straightforward to show, using symmetry considerations and the fact that $\beta \in A_{k, a_p}$ [so that $(\#\beta_i \neq 0) = k$], that the best such estimator is

$$\hat{\beta}_i^*(I_1', \ldots, I_p', J) = (1 - J)\left[I_i' a_p k \Big/ \sum_{i=2}^{p} I_i'\right] + J I_i' a_p.$$

Thus, for any estimator $\hat{\beta}(Y_2, \ldots, Y_p)$,

$$\max_{\beta \in A_{k, a_p}} R(\beta, \hat{\beta}) \geq \max_{\beta \in A_{k, a_p}} R(\beta, \hat{\beta}^*). \tag{A.5}$$

We now proceed to obtain a lower bound for the right-hand side of (A.5). Let $M_p \equiv (\sum_{i=2}^{p} I_i) - k$ be the number of Y_i's with $I_i = 1$ and $\beta_i = 0$. Note that M_p is Binomial $[\phi(\alpha_p)\Phi(\alpha_p - a_p)/\phi(\alpha_p - a_p), p - k - 1]$. Thus, for $\beta \in A_{k, a_p}$,

$$\begin{aligned} R(\beta, \hat{\beta}^*) = E_\beta |\hat{\beta}^* - \beta|^2 &= E_\beta |\hat{\beta}^* - \beta|^2 I[J = 0] + E_\beta |\hat{\beta}^* - \beta|^2 I[J = 1] \\ &= E_\beta \sum_{i=1}^{p} \left[\left[I_i' a_p k \Big/ \sum_{i=2}^{p} I_i'\right] - \beta_i\right]^2 I[J = 0] \\ &= E_\beta \sum_{I_i' = 1} \left[\left[a_p k \Big/ \sum_{i=2}^{p} I_i'\right] - \beta_i\right]^2 I[J = 0] \\ &= E_\beta \sum_{I_i = 1} \left[[a_p k/(M_p + k)] - \beta_i\right]^2 I[J = 0] \\ &= a_p^2 E_\beta \left[k\left[M_p/(M_p + k)\right]^2 + M_p\left[k/(M_p + k)\right]^2\right] I[J = 0] \\ &= a_p^2 k E_\beta \left[M_p/(M_p + k)\right] I[J = 0] \\ &= a_p^2 k \left[E_\beta\left[M_p/(M_p + k)\right] - E_\beta\left[M_p/(M_p + k)\right] I[J = 1]\right] \\ &\geq a_p^2 k \left[1 - E_\beta\left[k/(M_p + k)\right] - P[J = 1]\right]. \end{aligned} \tag{A.6}$$

Now let

$$a_p = \sqrt{2 \log p} - \log\log p \quad \text{and} \quad \alpha_p = a_p + (1/2)\log\log p. \tag{A.7}$$

For these choices, $\alpha_p - a_p \to \infty$ which yields $P[J = 1] \to 1$, and

$$\mathrm{E}[M_p] = (p - k - 1)\phi(\alpha_p)\Phi(\alpha_p - a_p)/\phi(\alpha_p - a_p) \to \infty,$$

which yields $E_\beta[k/(M_p + k)] \to 0$. Thus, for a_p and α_p in (A.7),

$$a_p^2 k\Big[1 - E_\beta\big[k/(M_p + k)\big] - P[J = 1]\Big] = 2k\log p - o(\log p). \tag{A.8}$$

Coupled with (A.6), (A.8) yields (A.4). □

The next result is the key to the proof of (i) in Theorem 5.1.

LEMMA A.3. *For* $\gamma_{\mathrm{LS}} \equiv (1, 1, \ldots, 1)$ *as in* (1.5) *and* γ_Π *as in* (3.1),

$$\big|X\widehat{\beta}_{\gamma_\Pi} - X\beta\big| \le (p\sigma^2\Pi)^{1/2} + \big|X\widehat{\beta}_{\gamma_{\mathrm{LS}}} - X\beta\big|. \tag{A.9}$$

PROOF. From the definition of γ_Π, it follows immediately that

$$\mathrm{SSE}_{\gamma_\Pi} + |\gamma_\Pi|\sigma^2\Pi \le \mathrm{SSE}_{\gamma_{\mathrm{LS}}} + p\sigma^2\Pi,$$

which implies

$$\big|X\widehat{\beta}_{\gamma_\Pi} - X\widehat{\beta}_{\gamma_{\mathrm{LS}}}\big|^2 \le p\sigma^2\Pi.$$

Inequality (A.9) now follows from the triangle inequality. □

The next result is the key to the proof of (ii) in Theorem 5.1.

LEMMA A.4. *Define* $W_k(a) \equiv E[\chi_k^2 \mid \chi_k^2 \ge \chi_{k,\,1-a}^2]$. *Then for* $\sigma^2 = 1$,

$$R\big(\beta, \widehat{\beta}_{\gamma_\Pi}\big) \le 2|\eta|(\Pi + 1) + 2\sum_\gamma \Big[2W_{|\gamma|}\big(P[\gamma_\Pi = \gamma]\big) - |\gamma|\Pi\Big]P[\gamma_\Pi = \gamma]. \tag{A.10}$$

PROOF. We begin by rewriting the risk of $\widehat{\beta}_{\gamma_\Pi}$ as

$$R\big(\beta, \widehat{\beta}_{\gamma_\Pi}\big) = E_\beta\big|X\widehat{\beta}_{\gamma_\Pi} - X\beta\big|^2 = \sum_\gamma E_\beta\Big(\big|X\widehat{\beta}_\gamma - X\beta\big|^2 I[\gamma_\Pi = \gamma]\Big).$$

We shall now bound each term $|X\widehat{\beta}_\gamma - X\beta|^2$. As in Figure A.1, define

$$a = \big|X\widehat{\beta}_{\gamma\eta} - X\widehat{\beta}_\gamma\big|, \qquad b = \big|X\widehat{\beta}_{\gamma\eta} - X\widehat{\beta}_\eta\big|, \qquad c = \big|X\widehat{\beta}_\eta - X\beta\big|,$$

where $X\widehat{\beta}_{\gamma\eta}$ is the projection of Y onto the space spanned by those X_i's specifie by γ and/or η. Using the obvious fact that $|X\widehat{\beta}_\gamma - X\beta| \le a + (b^2 + c^2)^{1/2}$ and the inequality $(x + y)^2 \le 2(x^2 + y^2)$, we have

$$|X\widehat{\beta}_\gamma - X\beta|^2 \le 2\big(a^2 + b^2 + c^2\big). \tag{A.11}$$

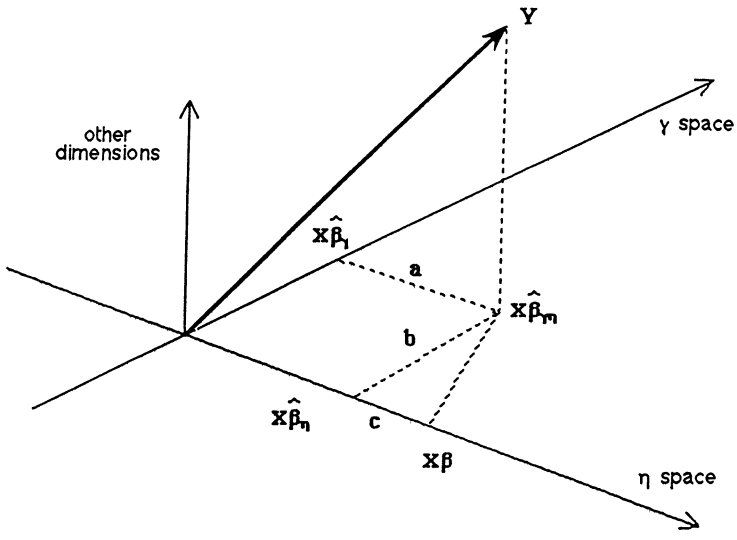

FIG. A.1.

Now from the definition of γ_Π, $(\mathrm{SSE}_\gamma + |\gamma|\Pi) \le (\mathrm{SSE}_\eta + |\eta|\Pi)$ on the set $[\gamma_\Pi = \gamma]$. Combined with the identities $\mathrm{SSE}_\gamma = a^2 + |Y - X\hat{\beta}_{\gamma\eta}|^2$ and $\mathrm{SSE}_\eta = b^2 + |Y - X\hat{\beta}_{\gamma\eta}|^2$, it follows that on $[\gamma_\Pi = \gamma]$,

$$a^2 + |\gamma|\Pi \le b^2 + |\eta|\Pi. \tag{A.12}$$

From (A.11) and (A.12), we have that on $[\gamma_\Pi = \gamma]$,

$$|X\hat{\beta}_\gamma - X\beta|^2 \le 2\big(2b^2 - |\gamma|\Pi\big) + 2\big(c^2 + |\eta|\Pi\big). \tag{A.13}$$

Now note that

$$\begin{aligned} E_\beta\big(b^2 I[\gamma_\Pi = \gamma]\big) &= E_\beta\big(b^2 \mid [\gamma_\Pi = \gamma]\big)P[\gamma_\Pi = \gamma] \\ &\le W_{|\gamma|}\big(P[\gamma_\Pi = \gamma]\big)P[\gamma_\Pi = \gamma] \end{aligned} \tag{A.14}$$

since for $k = |\gamma\eta| - |\eta|$,

$$W_k(a) = \sup_{A:P(A)=a}\Big\{E\big[\chi_k^2 \mid A\big]\Big\} = \sup_{A:P(A)=a}\Big\{E_\beta\big[|X\hat{\beta}_{\gamma\eta} - X\hat{\beta}_\eta|^2 \mid A\big]\Big\}$$

and for $i \le j$, $W_i(a) \le W_j(a)$. Combining (A.13) and (A.14) yields

$$\begin{aligned} E_\beta\big(|X\hat{\beta}_\gamma - X\beta|^2 I[\gamma_\Pi = \gamma]\big) \le{}& 2\Big[2W_{|\gamma|}\big(P[\gamma_\Pi = \gamma]\big) - |\gamma|\Pi\Big]P[\gamma_\Pi = \gamma] \\ &+ 2E_\beta\big(c^2 I[\gamma_\Pi = \gamma]\big) + 2|\eta|\Pi P[\gamma_\Pi = \gamma]. \end{aligned} \tag{A.15}$$

Summing (A.15) over γ and using the fact that $E_\beta c^2 = |\eta|$ yields (A.10). □

THEOREM 5.1. *Define* $\xi = \sqrt{\Pi}e^{(1-\Pi)/2}$. *For any* γ_Π,

(i) $$R(\beta, \hat{\beta}_{\gamma_\Pi}) \le p\sigma^2(\sqrt{\Pi}+1)^2.$$

(ii) $$R(\beta, \hat{\beta}_{\gamma_\Pi}) \le 2\sigma^2|\eta|(\Pi+1) + \sigma^2 4\sqrt{2}\big(\Pi/(\Pi-1)\big)^2 e^{p\xi}\sqrt{p\xi}.$$

PROOF. To prove (i), insert (A.9) of Lemma A.3 into $R(\beta, \hat{\beta}_{\gamma_\Pi}) = E_\beta|X\hat{\beta}_{\gamma_\Pi} - X\beta|^2$; (i) then follows directly.

To prove (ii), we consider the case $\sigma^2 = 1$ for simplicity. The extension for general σ^2 is straightforward. From (A.10) of Lemma A.4, we have

$$R(\beta, \hat{\beta}_{\gamma_\Pi}) \le 2|\eta|(\Pi+1) + 2\sum_{k=1}^{p}\binom{p}{k}\max_a a\big(W_k(a) - k\Pi\big). \tag{A.16}$$

Bounding the term inside the summation yields

$$\begin{aligned}
\max_a a\big(W_k(a) - k\Pi\big) &= \max_a \int_{\chi^2_{k,\,1-a}}^{\infty} (y - k\Pi)\, dP\big[\chi^2_k \le y\big] \\
&= \int_{k\Pi}^{\infty} (y - k\Pi)\, dP\big[\chi^2_k \le y\big] \\
&= C_k \int_{k\Pi}^{\infty} (y - k\Pi) e^{-y/2 + (k/2-1)\log y}\, dy \\
&\le C_k e^{-k\Pi/2 + (k/2-1)\log k\Pi} \\
&\quad \times \int_{k\Pi}^{\infty} (y - k\Pi) e^{-(y-k\Pi)(k\Pi - k + 2)/2k\Pi}\, dy \\
&= C_k e^{-k\Pi/2}(k\Pi)^{k/2-1}\big[2k\Pi/(k\Pi - k + 2)\big]^2 \\
&\le \big[2\Pi/(\Pi-1)\big]^2 C_k e^{-k\Pi/2}(k\Pi)^{k/2},
\end{aligned} \tag{A.17}$$

where $C_k = 2^{-k/2}/\Gamma(k/2)$, and the first inequality in (A.17) made use of the fact that $-y/2 + (k/2-1)\log y \le -k\Pi/2 + (k/2-1)\log k\Pi - (y - k\Pi)(k\Pi - k + 2)/2k\Pi$. Putting (A.17) into (A.16) yields

$$R(\beta, \hat{\beta}_{\gamma_\Pi}) \le 2|\eta|(\Pi+1) + 8\big[\Pi/(\Pi-1)\big]^2\left[\sum_{k=1}^{p}\binom{p}{k} C_k e^{-k\Pi/2}(k\Pi)^{k/2}\right]. \tag{A.18}$$

Using Stirling's formula, it can be shown that $C_k k^{k/2} \le \sqrt{k/2}e^{k/2}$. Letting

$\xi = \sqrt{\Pi}e^{(1-\Pi)/2}$ (for notational convenience) and bounding the summation in (A.18) yields

$$\begin{aligned}
&\sum_{k=1}^{p}\binom{p}{k}C_k e^{-k\Pi/2}(k\Pi)^{k/2} \\
&\quad\le \sum_{k=1}^{p}\binom{p}{k}\sqrt{k/2}\xi^k \\
&\quad= (1+\xi)^p\sum_{k=0}^{p}\binom{p}{k}\sqrt{k/2}\left[\frac{\xi}{1+\xi}\right]^k\left[\frac{1}{1+\xi}\right]^{p-k} \\
&\quad\le (1+\xi)^p\sqrt{p\xi/2(1+\xi)} \le e^{p\xi}\sqrt{p\xi/2(1+\xi)} \le e^{p\xi}\sqrt{p\xi/2}.
\end{aligned} \tag{A.19}$$

Inserting (A.19) into (A.18) yields

$$R(\beta, \hat{\beta}_{\gamma_\Pi}) \le 2|\eta|(\Pi+1) + 4\sqrt{2}\left[\Pi/(\Pi-1)\right]^2 e^{p\xi}\sqrt{p\xi},$$

which is exactly the conclusion (ii). □

The following result is the basis for the proof of Theorem 5.3. For the sake of brevity, the proof is only sketched, leaving the details to the reader.

LEMMA A.5. *For each* Π, *there exist* X *and* β *such that*

$$R(\beta, \hat{\beta}_{\gamma_\Pi}) \ge \sigma^2\left[(|\eta|-1)4\log p - o(\log p)\right]. \tag{A.20}$$

PROOF. For $\Pi < 2\log p - 2\log\log p$, it suffices to consider the case where X is orthogonal and appeal to (4.7) and (4.8).

For $\Pi \ge 2\log p - 2\log\log p$, the bound (A.20) can be obtained by constructing a "maliciously collinear" X of the form $X = [X_0, X^1, X^2, \ldots, X^M]$, where $X_0 = (1, \ldots, 1)'$, each $X^j = [X_1^j, \ldots, X_m^j]$ is $n \times m$ and $X_0, X^1, \ldots, X^M$ are orthogonal. (Since X is $n \times p$, $Mm + 1 = p$.) For this X, consider β of the form $\beta = (0, e, \ldots, e)'$, where $e = (0, \ldots, 0, 1)$. Thus, only the M coefficients of $X_m^1, \ldots, X_m^M$ are nonzero.

To obtain poor performance with this setup, the idea is to distribute $X_1^j, \ldots, X_{m-1}^j$ about X_m^j for each j, so that with high probability γ_Π substitutes an "incorrect" X_i^j for X_m^j, and the bias from such a substitution is substantial. This is obtained when, for each j, $X_1^j, \ldots, X_{m-1}^j$ are uniformly distributed on the surface of a d-dimensional sphere, $d = 2\log p - (\log p)^{3/4}$, of radius $r = \sqrt{2d}$ which is centered at X_m^j and is orthogonal to X_m^j. It can then be shown that for $|X_m^j|$ large enough, the probability of an "incorrect" substitution for X_m^j is $\approx 1 - (1 - \Phi(-\sqrt{d}))^{p/M} \to 1$ as $p \to \infty$ (for M fixed). This yields $R(\beta, \hat{\beta}_{\gamma_\Pi}) \approx \sigma^2 M r^2$ which in turn yields (A.20). □

The following result is required for the proof of Theorem 6.1 which is given afterward.

LEMMA A.6. *For any fixed* Π_0 *and* $\varepsilon > 0$, δ *can be chosen such that for* $|\Pi - \Pi_0| < \delta$,

$$\sup_{\beta} \left| R(\beta, \hat{\beta}_{\gamma_\Pi}) - R(\beta, \hat{\beta}_{\gamma_{\Pi_0}}) \right| < \varepsilon.$$

PROOF. Define $A_\delta = \{\Pi: |\Pi - \Pi_0| < \delta\}$ and $A_\delta^* = \{Y: \gamma_\Pi \neq \gamma_{\Pi_0}$ for some $\Pi \in A_\delta\}$. Now for any $\Pi \in A_\delta$,

$$\begin{aligned} \left| R(\beta, \hat{\beta}_{\gamma_\Pi}) - R(\beta, \hat{\beta}_{\gamma_{\Pi_0}}) \right| &= \left| E_\beta \left[\left| X\hat{\beta}_{\gamma_\Pi} - X\beta \right|^2 - \left| X\hat{\beta}_{\gamma_{\Pi_0}} - X\beta \right|^2 \right] \right| \\ &\leq \left| E_\beta \left[\left| X\hat{\beta}_{\gamma_\Pi} - X\beta \right|^2 - \left| X\hat{\beta}_{\gamma_{\Pi_0}} - X\beta \right|^2 \right] I[Y \in A_\delta^*] \right| \\ &\leq E_\beta W I[Y \in A_\delta^*], \end{aligned}$$

where $W = (\sigma \sqrt{2p(\Pi_0 + \delta)} + |X\hat{\beta}_{\gamma_{\mathrm{LS}}} - X\beta|^2)^2$. This last inequality is obtained from (A.9) of Lemma A.3. Because W does not depend on β and has finite mean, it suffices to show that

$$\sup_{\beta} P_\beta[Y \in A_\delta^*] \to 0 \quad \text{as } \delta \to 0. \tag{A.21}$$

Define

$$S_{\gamma, \gamma', \Pi} = \begin{cases} \{Y: \mathrm{SSE}_\gamma + |\gamma|\sigma^2\Pi = \mathrm{SSE}_{\gamma'} + |\gamma'|\sigma^2\Pi\}, & \text{if } X\hat{\beta}_\gamma \neq X\hat{\beta}_{\gamma'} \\ \emptyset, & \text{otherwise.} \end{cases}$$

Because

$$A_\delta^* \subset \bigcup_{\gamma, \gamma'} \bigcup_{\Pi \in A_\delta} S_{\gamma, \gamma', \Pi},$$

it suffices to show that for each γ, γ',

$$\sup_{\beta} P_\beta \left[\bigcup_{\Pi \in A_\delta} S_{\gamma, \gamma', \Pi} \right] \to 0 \quad \text{as } \delta \to 0. \tag{A.22}$$

Each nonempty $S_{\gamma, \gamma', m\,\Pi}$ may be expressed as $S_{\gamma, \gamma'\,\Pi} = \{Y: Y'M_{\gamma, \gamma'}Y = \Pi\}$ where $M_{\gamma, \gamma'}$ is a real symmetric matrix, so that

$$\bigcup_{\Pi \in A_\delta} S_{\gamma, \gamma', \Pi} = \{Y: Y'M_{\gamma, \gamma'}Y \in A_\delta\}.$$

Based on a diagonal decomposition of $M_{\gamma,\gamma'}$, we may write $Y'M_{\gamma,\gamma'}Y = \Sigma_{i=1}^{n}\lambda_i Z_i^2$ where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of $M_{\gamma,\gamma'}$ and $Z_1, \ldots, Z_n$ are independent normal random variables with possibly different means depending on β. [$(Z_1, \ldots, Z_n)' = UY$ for some orthonormal U.] Assuming (wlog) that $\lambda_1 \neq 0$,

$$\begin{aligned}\sup_\beta P_\beta\left[\bigcup_{\Pi\in A_\delta} S_{\gamma,\gamma'\,\Pi}\right] &= \sup_\beta P_\beta[Y'M_{\gamma,\gamma'}Y \in A_\delta] \\ &= \sup_\beta P_\beta\left[\sum_{i=1}^{n}\lambda_1 Z_i^2 \in A_\delta\right] \\ &\le \sup_\beta \sup_a P_\beta\left[\lambda_1 Z_1^2 \in [a-\delta, a+\delta]\right].\end{aligned} \tag{A.23}$$

Using standard methods, it is straightforward to show that the last term in (A.23) goes to 0 as $\delta \to 0$. This shows (A.22) which in turn shows (A.21). □

THEOREM 6.1. *Suppose that $\widehat{\Pi}_k$ is independent of $\widehat{\beta}_{\gamma_{\mathrm{LS}}}$ and that for some Π_0, $E|\widehat{\Pi}_k - \Pi_0| \to 0$ as $k \to \infty$. Then*

(i) $$\sup_\beta \left|R\left(\beta, \widehat{\beta}_{\gamma_{\widehat{\Pi}_k}}\right) - R\left(\beta, \widehat{\beta}_{\gamma_{\Pi_0}}\right)\right| \to 0;$$

(ii) $$\mathrm{RI}\left(\gamma_{\widehat{\Pi}_k}\right) \to \mathrm{RI}(\gamma_{\Pi_0}).$$

PROOF. For $\delta > 0$, define $A = \{\Pi\colon |\Pi - \Pi_0| \le \delta\}$. Let G_k be the probability distribution of $\widehat{\Pi}_k$. Because of the independence assumption on $\widehat{\Pi}_k$,

$$\begin{aligned}&\sup_\beta \left|R\left(\beta, \widehat{\beta}_{\gamma_{\widehat{\Pi}_k}}\right) - R\left(\beta, \widehat{\beta}_{\gamma_{\Pi_0}}\right)\right| \\ &\quad= \sup_\beta \left|\int_0^\infty R\left(\beta, \widehat{\beta}_{\gamma_\Pi}\right) - R\left(\beta, \widehat{\beta}_{\gamma_{\Pi_0}}\right) dG_k(\Pi)\right| \\ &\quad\le \sup_\beta \int_0^\beta \left|R\left(\beta, \widehat{\beta}_{\gamma_\Pi}\right) - R\left(\beta, \widehat{\beta}_{\gamma_{\Pi_0}}\right)\right| dG_k(\Pi) \\ &\quad\le \sup_\beta \left[\int_A \left|R\left(\beta, \widehat{\beta}_{\gamma_\Pi}\right) - R\left(\beta, \widehat{\beta}_{\gamma_{\Pi_0}}\right)\right| dG_k(\Pi)\right. \\ &\qquad\left. + \int_{\bar{A}} R\left(\beta, \widehat{\beta}_{\gamma_\Pi}\right) dG_k(\Pi) + \int_{\bar{A}} R\left(\beta, \widehat{\beta}_{\gamma_{\Pi_0}}\right) dG_k(\Pi)\right] \\ &\quad\le \int_A \sup_\beta \left|R\left(\beta, \widehat{\beta}_{\gamma_\Pi}\right) - R\left(\beta, \widehat{\beta}_{\gamma_{\Pi_0}}\right)\right| dG_k(\Pi) \\ &\qquad + \sup_\beta \int_{\bar{A}} R\left(\beta, \widehat{\beta}_{\gamma_\Pi}\right) dG_k(\Pi) + R\left(\beta, \widehat{\beta}_{\gamma_{\Pi_0}}\right) P(\widehat{\Pi}_k \in \bar{A}).\end{aligned} \tag{A.24}$$

We shall now show that for any $\varepsilon > 0$, δ and K can be chosen so that for $k \geq K$, the sum of these last three expressions is less than ε. By Lemma A.6, δ can be chosen small enough so that $\sup_\beta |R(\beta, \widehat{\beta}_{\gamma_\Pi}) - R(\beta, \widehat{\beta}_{\gamma_{\Pi_0}})| \leq \varepsilon/3$ on A. Thus for all k,

$$\text{(A.25)} \quad \int_A \sup_\beta \left|R\left(\beta, \widehat{\beta}_{\gamma_\Pi}\right) - R\left(\beta, \widehat{\beta}_{\gamma_{\Pi_0}}\right)\right| dG_k(\Pi) \leq (\varepsilon/3) P\left(\widehat{\Pi}_k \in A\right) \leq \varepsilon/3.$$

For the second expression, note that by Theorem 5.1, we may select c_1 and c_2 such that $R(\beta, \widehat{\beta}_{\gamma_\Pi}) < c_1\Pi + c_2$ for some c_1 and c_2. By the convergence of $\widehat{\Pi}_k$, we can choose K' so that for $k \geq K'$,

$$\begin{aligned} &\sup_\beta \int_{\bar{A}} R\left(\beta, \widehat{\beta}_{\gamma_\Pi}\right) dG_k(\Pi) \\ &\quad \leq \int_{\bar{A}} (c_1\Pi + c_2)\, dG_k(\Pi) \\ &\quad \leq c_1 \int_{\bar{A}} |\Pi - \Pi_0|\, dG_k(\Pi) + (c_1\Pi_0 + c_2) P(\widehat{\Pi}_k \in \overline{A}) \\ &\quad \leq c_1 E|\widehat{\Pi}_k - \Pi_0| + (c_1\Pi_0 + c_2) P(\widehat{\Pi}_k \in \bar{A}) < \varepsilon/3. \end{aligned} \tag{A.26}$$

For the third term, again because of convergence, we can choose K'' so that for $k \geq K''$,

$$R\left(\beta, \widehat{\beta}_{\gamma_{\Pi_0}}\right) P(\widehat{\Pi}_k \in \overline{A}) < \varepsilon/3. \tag{A.27}$$

Finally, combining (A.24)–(A.27), it follows that for $k \geq K \equiv \max\{K', K''\}$,

$$\sup_\beta \left|R\left(\beta, \widehat{\beta}_{\gamma_{\widehat{\Pi}_k}}\right) - R\left(\beta, \widehat{\beta}_{\gamma_{\Pi_0}}\right)\right| < \varepsilon,$$

which proves (i). Conclusion (ii) is immediate from this and the definition of RI. □

REFERENCES

AKAIKE, H. (1970). Statistical predictor identification. *Ann. Inst. Statist. Math.* **22** 203–217.

AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control* **19** 716–723.

ALLEN, D. M. (1971). Mean square error of prediction as a criterion for selecting variables. *Technometrics* **13** 469–475.

ALLEN, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics* **16** 125–127.

COHEN, A. (1965). A hybrid problem on the exponential family. *Ann. Math. Statist.* **36** 1185–1206.

DONOHO, D. L. and JOHNSTONE, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–456.

HANNAN, E. J. and QUINN, B. G. (1979). The determination of the order of an autoregression. *J. Roy. Statist. Soc. Ser. B* **41** 190–195.

HOCKING, R. R. (1976). The analysis and selection of variables in linear regression. *Biometrics* **32** 1–49.

MALLOWS, C. L. (1973). Some comments on C_p. *Technometrics* **15** 661–676.
MILLER, A. J. (1990). *Subset Selection in Regression.* Chapman and Hall, New York.
RISSANEN, J. (1986). A predictive least squares principle. *IMA J. Math. Control Inform.* **3** 211–222.
SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464.
SHIBATA, R. (1981). An optimal selection of regression variables. *Biometrika* **68** 45–54.
STONE, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. Roy. Statist. Soc. Ser. B* **39** 44–47.
THEIL, H. (1961). *Economic Forecasts and Policy.* North-Holland, Amsterdam.
THOMPSON, M. L. (1978a). Selection of variables in multiple regression. A review and evaluation. *Internat. Statist. Rev.* **46** 1–19.
THOMPSON, M. L. (1978b). Selection of variables in multiple regression. II. Chosen procedures, computations and examples. *Internat. Statist. Rev.* **46** 129–146.
VENTER, J. H. and STEEL, S. J. (1992). Some contribution to selection and estimation in the normal linear model. *Ann. Inst. Statist. Math.* **44** 281–297.
WEI, C. Z. (1992). On predictive least squares principles. *Ann. Statist.* **20** 1–42.
ZELLNER, A. (1971). *An Introduction to Bayesian Inference in Econometrics.* Wiley, New York.

DEPARTMENT OF STATISTICS
UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA, PENNSYLVANIA 19104-6302

DEPARTMENT OF MANAGEMENT SCIENCE
AND INFORMATION SYSTEMS
UNIVERSITY OF TEXAS
AUSTIN, TEXAS 78712-1175