American Society for Quality

More Comments on CP
Author(s): Colin L. Mallows
Reviewed work(s):
Source: *Technometrics*, Vol. 37, No. 4 (Nov., 1995), pp. 362-372
Published by: American Statistical Association and American Society for Quality
Stable URL: http://www.jstor.org/stable/1269729
Accessed: 06/02/2013 16:26

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at
http://www.jstor.org/page/info/about/policies/terms.jsp

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

American Statistical Association and *American Society for Quality* are collaborating with JSTOR to digitize, preserve and extend access to *Technometrics.*

http://www.jstor.org

This content downloaded on Wed, 6 Feb 2013 16:26:47 PM
All use subject to JSTOR Terms and Conditions

© 1995 American Statistical Association and
the American Society for Quality Control

TECHNOMETRICS, NOVEMBER 1995, VOL. 37, NO. 4

More Comments on C_P

Colin L. MALLOWS

AT&T Bell Labs
Murray Hill, NJ 07974

I study the typical configuration of a Cp plot when the number of variables in the regression problem is large and there are many weak effects. I show that a particular configuration that is very commonly seen can arise in a simple way. I give a formula by means of which the risk incurred by the "minimum C_P" rule can be estimated.

KEY WORDS: Predictive mean squared error; Regression; Subset selection.

1. INTRODUCTION

Consider the problem of choosing a prediction formula, in the standard linear regression situation with many independent variables. In the usual model, one has

$$y = \eta + e, \tag{1}$$

where the elements of the $n \times 1$ vector y are the dependent observations and the deviations e are assumed to be independent with mean 0 and common variance σ^2. One also may assume that

$$\eta = X\beta, \tag{2}$$

where the columns of the $n \times k$ matrix X are the regressors. Consider the case in which k is large. A procedure that has received much attention and is commonly used is to select a subset of $K = \{1, 2, \ldots, k\}$ somehow and to fit the corresponding subset of terms by least squares. Let P be any subset of K, with $|P| = p$ members, and let X_P be the corresponding submatrix of X, containing only those columns whose indices are in P. Suppose that we fit the model $y = X_P\beta_P + e$ by least squares, obtaining a vector of estimated coefficients $\hat{\beta}_P$. Define the Predictive Error for this estimate to be $\mathrm{PE}_P = |\eta - X_P\hat{\beta}_P|^2$. We would like to choose P so that PE_P is small. Because η is not known, values of PE_P cannot be computed directly; in an attempt to deal with this difficulty, the statistic C_P is defined to be (Mallows 1973)

$$C_P = \frac{\mathrm{RSS}_P}{\hat{\sigma}^2} - n + 2p,$$

where $\mathrm{RSS}_P = |y - X_P\hat{\beta}_P|^2$ is the sum of squared residuals from the least squares fit of the subset P and $\hat{\sigma}^2$ is an unbiased estimate of σ^2, usually obtained as $\mathrm{RSS}_K/(n-k)$. Assumption (2) is needed only to make this estimate valid. If an alternative unbiased estimate of σ^2 is used (for example, based on replicates) then (2) need not hold.

C_P has the property that [if the model (1) is correct and the estimate of σ^2 is unbiased] $E(\hat{\sigma}^2 C_P) = E(\mathrm{PE}_P)$. This result holds, however, only if the subset P is chosen independently of the data in hand. The C_P plot (values of C_P against p) is a graphical device that facilitates examination of the C_P values corresponding to all 2^k possible subsets. It should be pointed out that, if $|\eta - X_P\beta_P|$ is large, then PE has a large variance and $E(\mathrm{PE})$ may not be a good approximation to PE. It is, of course, always important to check that no important structure has been ignored.

It is tempting to use a C_P plot to identify subsets with small values of C_P and hence to select a particular subset of regressors, but the expected PE of a (select, estimate by least squares) procedure is not estimated accurately by the C_P value associated with the selected minimizing subset because allowance has not been made for the fact that the selected subset may depend on the observed data. Mallows (1973) pointed out that the $E(\mathrm{PE})$ of the "fit the minimum-C_P subset by least squares" procedure can be much larger than that of the "include all regressors" procedure. It will be shown that in some cases this $E(\mathrm{PE})$ can be estimated directly from the C_P plot.

A very common configuration on a C_P plot is that seen in Figure 1. This is the C_P plot for the "evaporation" data given by Freund (1979) and included by Becker, Chambers, and Wilks (1988) in the data sets distributed with the S system. The response is the amount of evaporation from the soil, measured on 46 consecutive days: there are 10 independent variables, of which 9 fall into three groups of three highly correlated variables. Many of the well-fitting subsets include just one regressor from each set. The C_P plot exhibits a convex lower boundary, with many subsets giving C_P values close to this boundary. The plot shows that the data is ambiguous, with no subset being strongly indicated as "best." We will see that such a configuration can arise from a very simple mechanism. We will find also that in such a case, picking the "min C_P" subset and fitting by least squares will not give a good prediction formula.

Let us always set $Q = K - P$, with $|Q| = q = k - p$. An alternative formula for C_P is

$$C_P = \frac{\mathrm{SS}_Q}{\hat{\sigma}^2} - k + 2p = \frac{\mathrm{SS}_Q}{\hat{\sigma}^2} + k - 2q, \tag{3}$$

where $\mathrm{SS}_Q = \mathrm{RSS}_P - \mathrm{RSS}_K$.

This content downloaded on Wed, 6 Feb 2013 16:26:47 PM
All use subject to JSTOR Terms and Conditions

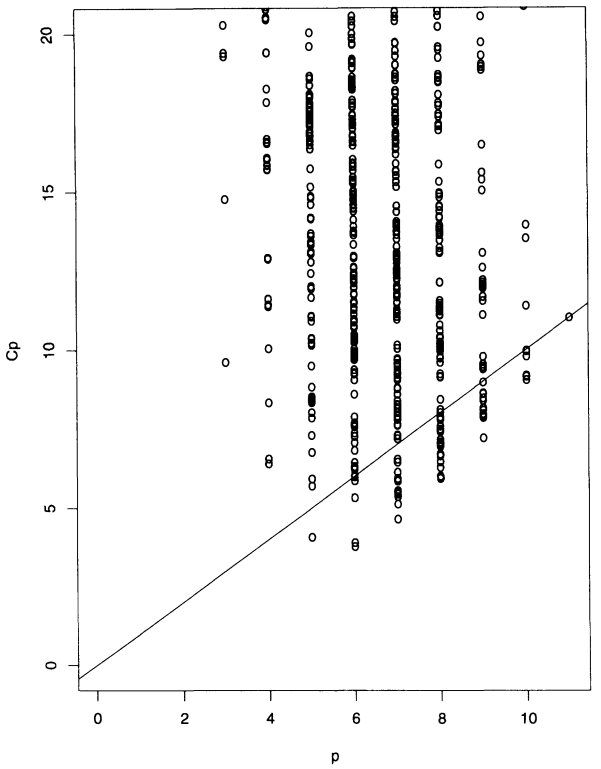

Figure 1. C_P Plot for the "Evaporation" Data (Becker, Chambers, and Wilks 1988). The response is the amount of evaporation from the soil, measured on 46 consecutive days. There are 10(+1) independent variables—maximum, minimum, and average soil temperatures (integrated area under daily soil temperature curve); maximum, minimum, and average air temperature; maximum, minimum, and average relative humidity; and total wind (miles per day). The average absolute correlation among these 10 variables is .47; the largest correlation is .95.

It is clear that, if the subset P^+ has $p+1$ members and contains the subset P, then $C_{P^+} - C_P = 2 - (\mathrm{SS}/\widehat{\sigma}^2)$, where SS is the 1-df sum of squares due to the $p+1$st variable. Now $\mathrm{SS}/\widehat{\sigma}^2$ is a t^2 statistic and is what a stepwise testing procedure (either forward or backward) will look at. The $p+1$st variable will be included at level α if $\mathrm{SS}/\widehat{\sigma}^2 > t_\alpha^2$, where t_α is the significance point of t with the appropriate number of degrees of freedom. Consider the subsets P that correspond to points on the lower envelope of the C_P plot, and among these consider a subset P_α that has the property that at some level α no additional variable can be included and no variable can be dropped. Then for each subset P' with $|p' - p| = 1$, we have $C_{P'} - C_{P_\alpha} > (2 - t_\alpha^2)(p' - p)$. Thus the slope of the lower boundary of the C_P plot is about $2 - t_\alpha^2$. More precisely, the tangent line with slope $2 - t_\alpha^2$ will pass through the point (p, C_{P_α}). The tangent with slope 0 (i.e., $t^2 = 2$) gives the "minimum C_P" subset. Note that in general the "lower boundary" subsets need not be nested; for example, it is possible that the best single variable is not a member of the best-fitting pair.

Several suggestions have been made that are equivalent to replacing the magic multiplier 2 in the C_P formula with something larger. Using a multiplier of m corresponds to setting $t_\alpha^2 = m$ and will determine the subset at which the tangent is about $2 - m$; thus, if $m = \log n$ (and n is larger than 7), this slope is negative, and the procedure will find a "best" subset with fewer variables than does "minimize C_P."

Let us search for an asymptotic theory that captures (to some degree) the behavior seen in Figure 1. Note that most previous asymptotic theories [e.g., Stone (1977) and Schwarz (1978); two exceptions were those of Shibata (1981) and Breiman and Freedman (1983)] have assumed that k remains fixed as $n \to \infty$ and that some of the variables have zero coefficients, but the rest have coefficients that are nonzero. In such a case the asymptotic configuration (as $n \to \infty$) of the C_P plot will be very different from Figure 1, being more like Figure 2, which was constructed as follows. I took the evaporation data and identified the best fitting subset with $p = 6$ (constant + five variables), which I call P_6. Let $\widehat{y}_6$ be values of "fit + residual" for this subset. Artificial data were then generated by replacing y by $(1/4)y + (3/4)\widehat{y}_6$. Thus, for these modified data the fit for the subset P_6 is unchanged, with the residuals being reduced by a factor of 4 (simulating the effect of multiplying n by 16), while for the fit of the complete model the magnitudes of the coefficients corresponding to variables not in P_6, and also the residuals, are shrunk

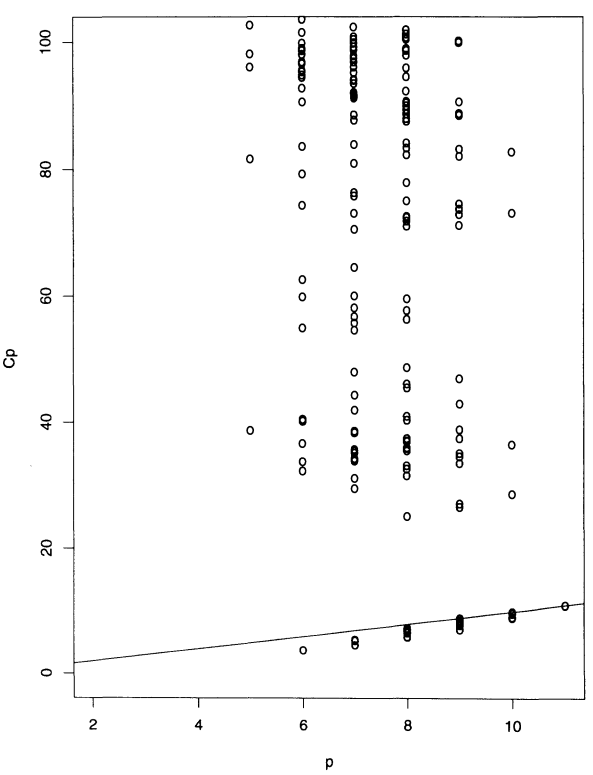

Figure 2. C_P Plot for the Modified "Evaporation" Data. See the text for explanation.

This content downloaded on Wed, 6 Feb 2013 16:26:47 PM
All use subject to JSTOR Terms and Conditions

by a factor of 4. We see that the resulting C_P plot is very different from Figure 1. There is a narrow cloud of points near the line $C_P = p$, each corresponding to a subset that includes P_6, but all subsets that do not contain P_6 give very large C_P values. For these data, all reasonable selection rules should identify the same six-variable subset, P_6. This will be the case for all of the stepwise rules that correspond to tangents with slopes between +.888 and −34.892—that is, corresponding to two-tailed tests with sizes .398 ($t^2 = 1.112$) and .0000004 ($t^2 = 36.892$), respectively. Clearly for these artificial data the choice of the slope is not very critical.

An asymptotic theory that keeps the regression coefficients constant as the number of observations increases of infinity will be dealing with configurations more like Figure 2 than Figure 1. In practice, such a configuration is very unusual; it can arise only when the data are unambiguous. It is much more commonly the case (as in Fig. 1) that no single subset is indicated clearly.

In the following sections, two versions of an asymptotic theory are developed in which interpretation of the data is ambiguous, even in the limiting case. This is achieved by allowing the number of variables k to increase without limit. In the first version of the theory, the regressors are orthogonal. A simple formula (8) provides an estimate of PE that will be incurred by using a prediction rule of the form "choose some tangent slope $2 - t^2$; select the corresponding extreme subset $P(t)$ and fit the coefficients by least squares." In the second version the regressors are not all orthogonal, but in one (important) case the same formula (8) for the PE continues to apply. Numerical study shows that in many cases the PE of any "subset–least squares" rule exceeds that of the rule "fit all coefficients by least squares." This fails to happen only when many coefficients are in fact very near 0 (relative to their standard errors), but it is very difficult to tell from the C_P plot whether or not this is so. The conclusion is that, no matter how the subset is chosen, "subset–least squares" cannot be relied on when the objective is prediction.

2. ORTHOGONAL REGRESSORS

Let us start by considering the simplest case, in which the $k = k(n)$ independent variables are uncorrelated and standardized and either (2) holds with $k \ll n$ so that there are plenty of degrees of freedom for estimating the residual variance, or [if (2) fails] an alternative consistent estimate of σ^2 is avaiable. In either case asymptotically we may assume that σ^2 is known, and (rescaling the design if necessary) that $\sigma^2 = 1$. Thus we have $X^T X = I_k$, and asymptotically

$$b_j = \beta_j + e_j, \qquad j = 1, 2, \ldots, k, \tag{4}$$

where $\beta = X^T \eta$ and the e's are iid standard Gaussian. From here on we may assume that $X^T = (I_k, 0(k \times n - k))$ so that $y_j = \beta_j + e_j$ for $j = 1, \ldots, k$, $y_j = e_j$ for $j = k + 1, \ldots, n$ with var(e) known = 1. Then $b_j = y_j$ for $j = 1, \ldots, k$.

We must specify how the β's are distributed as $k \to \infty$. What follows is a hand-waving argument; rigorous proofs will have to wait for another occasion. First, we need not specify the values of the "large" β's, where "large" means that the chance of mistaking such a β for 0 is vanishingly small. Something must be assumed about the "small" coefficients, and this assumption will not be innocuous. Suppose that some number k' of the β's are not "large"; assume that as $k \to \infty$ (and $k' \to \infty$), the empirical distribution of these "small" β's converges to a limit G; that is, assume that, for all h,

$$\frac{1}{k'} \sum_j [\beta_j < h] \to G(h) \tag{5}$$

for some function G. (Here the notation $[A]$ means 1 or 0 according as the logical statement A is true or false.) An interesting special case of (5) is

$$G(h) = \frac{1}{2}\left(1 + \frac{h}{\tau}\right), \qquad -\tau < h < \tau, \tag{6}$$

where τ is large (relative to the common standard error of the b's, which is unity). In this case the β's are locally uniform (i.e., near the origin relative to their standard errors), with density $\lambda = k'/2\tau$ per unit. We will see that essentially the same results are obtained if we assume that the β's are Gaussian with mean 0 and (large) scale τ', if $\lambda = k'/\tau'\sqrt{2\pi}$.

Appendix A shows that under these assumptions several things are true. First, in the special case (6) the shape of the lower boundary of the C_P plot is always the same, except for rescaling by λ in both directions, and is a cubic polynomial

$$C_P \approx k + \frac{q^3}{12\lambda^2} - 2q, \tag{7}$$

and

$$\frac{1}{\lambda}(C_P - k) = \frac{1}{12}\left(\frac{q}{\lambda}\right)^3 - 2\left(\frac{q}{\lambda}\right).$$

See Figure 3. In this limiting case, C_P is a minimum at $q = 2\sqrt{2}\lambda$ corresponding to two-tailed tests with critical value $t = \sqrt{2}$, size 15.73%. Moreover, the lower boundary meets the line $C_P = p$, where $q = 2\sqrt{3}\lambda$, corresponding to two-tailed tests with critical value $t = \sqrt{3}$, size 8.3%. The boundary reaches the height $C_P = k$, where $q = 2\sqrt{6}\lambda$, corresponding to two-tailed tests with $t = \sqrt{6}$, size 1.4%.

Second, let $P(t)$ be the subset that is determined by t tests with critical value t—that is, $P(t) = \{j: |b_j| > t\}$—and let $Q(t)$ be the complementary subset. Then $C_{P(t)}$ lies on the lower boundary of the C_P plot, and the slope of the boundary there is $2 - t^2$. Let $\mathrm{PE}(t) = E(\mathrm{PE}_{P(t)})$ be the expected summed prediction error for this subset, which in this orthogonal case reduces to

$$|\eta - X\beta|^2 + \sum_{Q(t)} \beta^2 + \sum_{P(t)} (\beta - b)^2.$$

This content downloaded on Wed, 6 Feb 2013 16:26:47 PM
All use subject to JSTOR Terms and Conditions

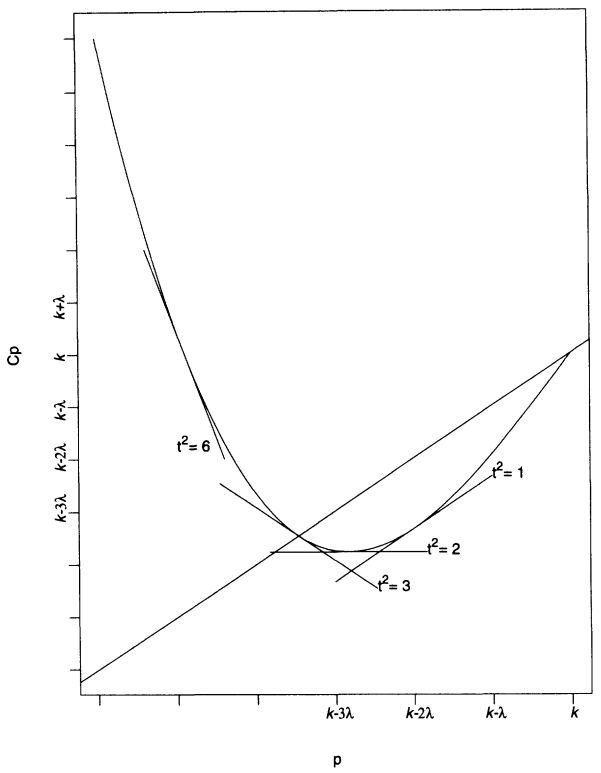

Figure 3. The Canonical Shape of the Lower Boundary of a C_P Plot in the "Asymptotically Uniform" Case. λ is the density of the true regression coefficients, near the origin. The tangent with slope $2 - t^2$ corresponds (asymptotically) to using t tests with critical value t.

The first term is not affected by the subset-selection procedure, and so can be ignored. Define

$$\mathrm{PE}_{P(t)} = \sum_{Q(t)} \beta^2 + \sum_{P(t)} (\beta - b)^2.$$

Appendix A shows that for all G there is asymptotically a relationship between $\mathrm{PE}_{P(t)}$ and the curvature of the lower boundary of the C_P plot; namely,

$$\mathrm{PE}(t) = C_{P(t)} + \frac{4t^2}{C''_{P(t)}} \tag{8}$$

$$= C_{P(t)} + \frac{4(2 - C'_{P(t)})}{C''_{P(t)}}. \tag{9}$$

Thus at the $\min C_P$ subset, we have $t = \sqrt{2}$, and the expected PE is not the observed value of C_P but rather

$$\min C_P + 8/C''_p. \tag{10}$$

In the "asymptotically uniform" case of (6), this becomes simply

$$\mathrm{PE} = \min C_P + 2q. \tag{11}$$

In Figure 1, a rough estimate is that, at the minimum, C'' is about 1.0 (to about 1.5 significant digits), so the estimate $E(\mathrm{PE})$ for the minimum C_P subset is about $3.8 + 8/1 = 11.8$. Using the "uniform" result (11) we get an estimate of PE of about $3.8 + 10 = 13.8$. Although these estimates are very rough indeed and are in any case irrelevant because in Figure 1 we do not have orthogonal regressors, they do suggest that the observed value of $\min C_P$ is a very poor estimate of the actual predictive error incurred by using the $\min C_P$ rule and that for these data little is to be gained by taking $|P| < k$. A subsequent section reports on a Monte Carlo study to explore how far off this estimate might be.

The effect of inaccuracy in $\hat{\sigma}$ is easily derived, at least in the "asymptotically uniform" case. Suppose that the model (4), (5) holds [with $\mathrm{var}(e) = 1$] but we have only an estimate $\hat{\sigma}^2$ of $\mathrm{var}(e)$. Then the minimizing value of C_P with $|P| = k - q$ is asymptotically

$$k + \frac{q^3}{12\lambda^2\hat{\sigma}^2} - 2q$$

so that the shape of the C_P plot is unchanged; the only effect is to replace λ by $\lambda\hat{\sigma}$. The expected PE of the boundary subset corresponding to slope $2 - t^2$ is asymptotically

$$E(\mathrm{PE}) = k + \frac{2}{3}\lambda t^3 E(\hat{\sigma}^3), \tag{12}$$

whereas Formulas (8) and (9) give the estimate

$$k + \frac{2}{3}\lambda t^3 \hat{\sigma}. \tag{13}$$

There appears to be no simple way to allow for the difference between (12) and (13). An indication of the magnitude of the difference is that, if $\hat{\sigma}^2$ is distributed as χ^2_ν/ν, then $E(\hat{\sigma}^3)/E(\hat{\sigma})$ is $(1 + 1/\nu)$.

Figures 4 and 5 show the effect of taking G to be, respectively, (a) a mixture with a spike at 0 of height k_0 and a slab (indefinitely wide) of height λ, for various values of k_0/λ and (b) Gaussian with scale τ, for various values of τ. Each figure shows several pairs of curves, one pair for each value of the parameter (k_0/λ or τ). The lower curve gives the shape of the lower boundary of the C_P plot, rescaled to make the minimum occur at the abscissa value $k - 2\sqrt{2}\lambda$. The upper curve gives the corresponding expected PE's. In Figure 4, the flattest C_P curve and the highest PE curve correspond to $k_0 = 0$; that is, no true coefficients are exactly 0. This C_P curve agrees with the one in Figure 3. The lowest C_P curve corresponds to the case $\lambda = 0$; that is, all nonlarge coefficients are exactly 0. Here the C_P boundary rises to meet the line $C_P = p$ at $p = k - 3.356\lambda$, corresponding to the subset that omits exactly all terms with zero coefficients. Here the PE is also p, because there is no bias and only the variance term contributes. All smaller subsets have very large C_P values (as in Fig. 2), and correspondingly very large PE values. Examination of the algebra shows that PE is monotone decreasing as a function of p whenever $k_0/\lambda < \sqrt{2\pi}$. Thus, in such cases the optimum subset has $p = k$. Moreover, all of the (p, C_P) curves in the figure have very similar shapes; in most applications it would be very difficult to infer an appropriate value of k_0/λ from the shape of the lower boundary of the (p, C_P) plot. In Figure 5, the flattest C_P curve and the highest PE curve both correspond

This content downloaded on Wed, 6 Feb 2013 16:26:47 PM
All use subject to JSTOR Terms and Conditions

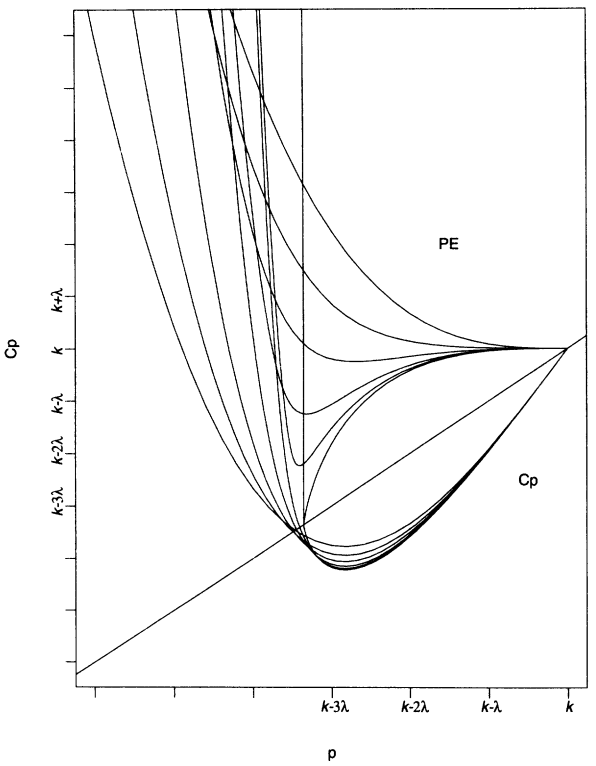

Figure 4. This Shows (p, min Cp) and (p, PE) Curves for Six "Orthogonal Regressors, Spike + Slab Prior" Specifications; Namely, $k_0/\lambda = 0, 2, 6, 18, 54, \infty$. In each case, the curves have been scaled so that the minimum C_P occurs at the same abscissa value. The top PE curve and the broadest C_P curve both correspond to the "infinity" case. The PE curve is monotone decreasing (and $>k$) whenever $k_0/\lambda < \sqrt{2\pi}$.

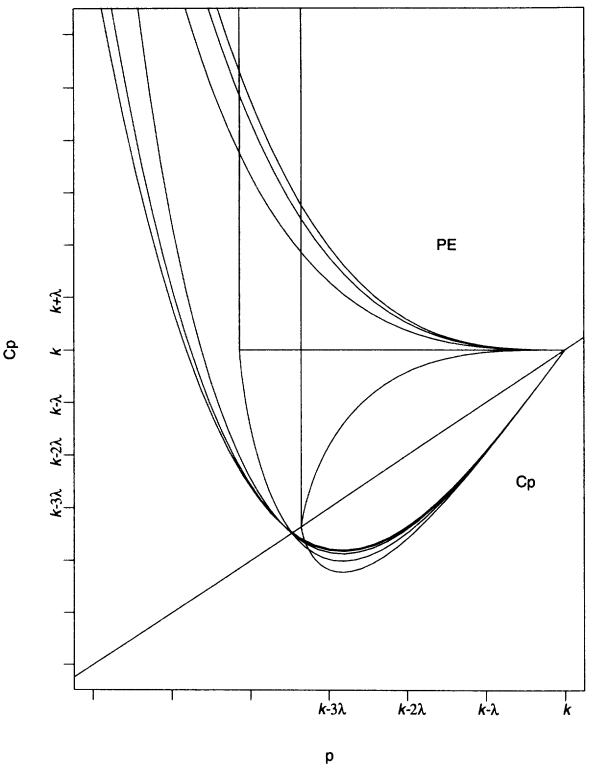

Figure 5. This Shows (p, min Cp) and (p, PE) curves for six "Orthogonal Regressors, Gaussian Prior" Specifications, With Scales $\tau = 0, 1, 2, 3, 4$, Respectively. In each case, the curves have been scaled so that the minimum C_P occurs at the same abscissa value. The highest PE curve and the broadest C_P curve both correspond to the case $\tau = 4$. A section of the PE curve is flat for $\tau = 1$.

to $\tau = 4$. The lowest C_P curve corresponds to $\tau = 0$, which reproduces the case $\lambda = 0$ of Figure 4. We see that both the C_P and PE curves are very similar to those in Figure 4 and deduce that in practice one will not be able to distinguish between the two cases in which (1) some coefficients are really 0 and (2) the true coefficients are distributed roughly Gaussianly.

3. CORRELATED REGRESSORS

The following formulation attempts to deal with the correlated case. I admit that I do not find it completely satisfying; to get a tractable result I have needed to build in a lot of independence. Moreover, I cannot go beyond the "asymptotically uniform coefficients" case.

Let X be any $n \times k$ matrix, with $n > k$, of rank k, and suppose that the design matrix $X^{(m)}$ is given by the Kronecker product

$$X^{(m)} = I_m \times X \tag{14}$$

so that $X^{(m)}$ is $mn \times mk$ and contains m copies of X down its diagonal. For simplicity, there is no additional constant term. m is going to go to infinity with X held fixed. There are $m(n-k)$ df for estimating σ^2, so asymptotically we may assume that σ^2 is known and equal to unity.

Assume that as $m \to \infty$, the empirical distribution of the β_j's approaches a Gaussian distribution with mean 0, and covariance $\tau^2 V$ where τ is large (relative to unity). Our results involve the ratio m/τ. As in the orthogonal case, the only important part of this assumption is the behavior of the limiting empirical distribution of the β_j's near the origin. We assume that this is locally uniform. Under these assumptions, Appendix B shows that the asymptotic shape of the C_P plot is the same as in the orthogonal case, and the Formulas (8) and (10) for the predictive error of the "boundary subset" rules still hold.

I have been unable to find similar explicit asymptotic results for the case of arbitrary $G(\beta)$, in the nonorthogonal case. Numerical simulations suggest, however, that (8) may give useful guidance quite generally. In Figures 6 and 7, I take $k = 2$, with the β's distributed according to a centered (circular) Gaussian with scale τ and with

$$X = \begin{pmatrix} c & s \\ s & c \end{pmatrix},$$

where $c^2 + s^2 = 1$. Thus, within each pair of variables we have a correlation of $2cs$. First consider the effect of decreasing c, so that the within-pair correlation increases from 0; specifically let $c = 1, .95, .9, .85, .8, .75$, corresponding to pairwise correlations of

This content downloaded on Wed, 6 Feb 2013 16:26:47 PM
All use subject to JSTOR Terms and Conditions

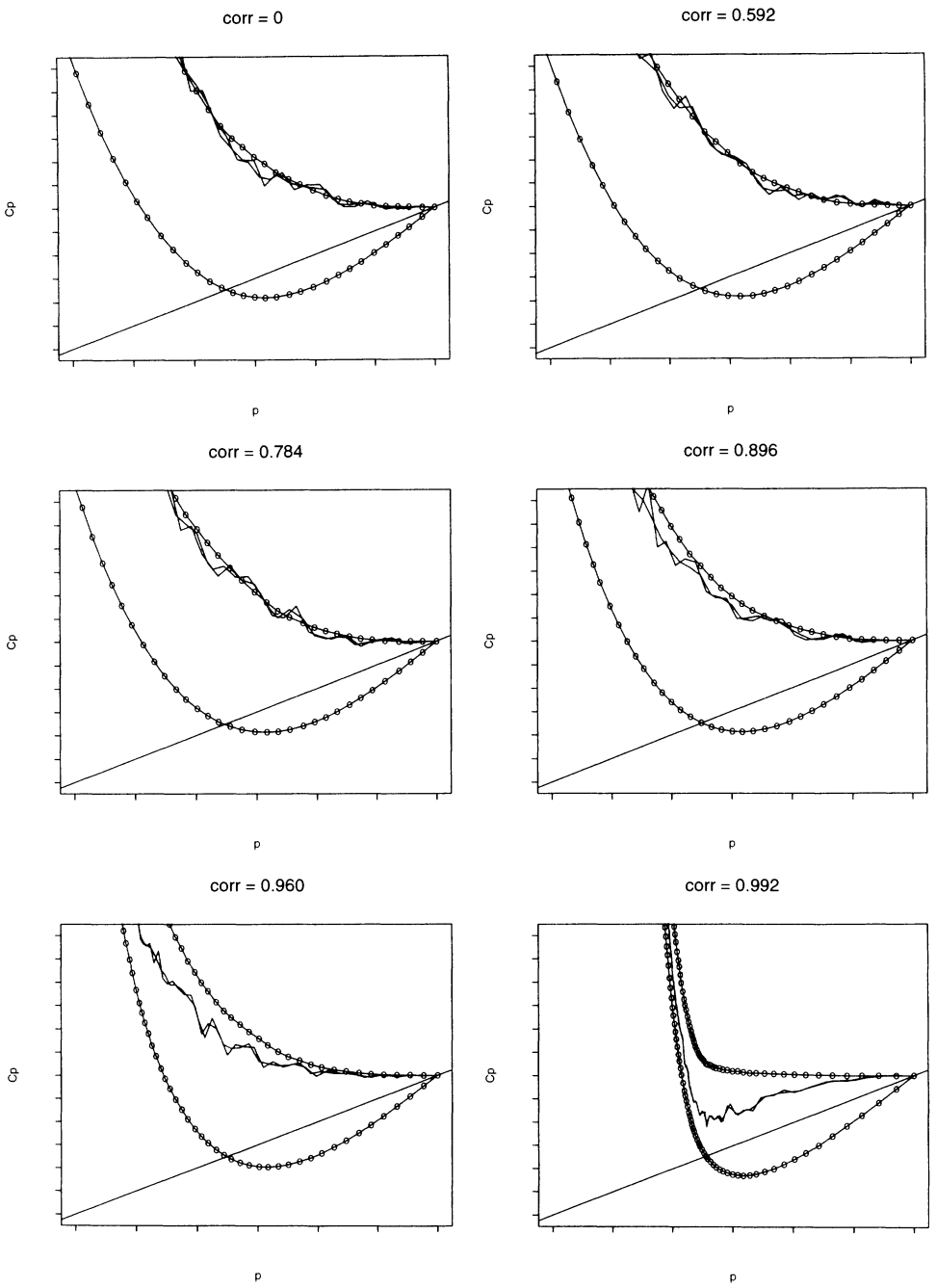

Figure 6. Monte Carlo Estimates of the Average C_P, the Corresponding PE, and Two Estimates of PE, for the Situation Described in the Text, With $k = 2$ and Various Values of the Within-Pair Correlation.

0, .592, .784, .896, .960, and .992, respectively. In this simulation, m is indefinitely large, and $\tau = 10$. Figure 6 shows the average (over 10,000 simulations) of the resulting C_P and PE values, computed for values of $h = 0(.1)6$, and also two versions of the estimated PE, corresponding to (8) and (9), respectively. We see that the average estimated PE is close to the average realized PE in all cases, except when the within-pair correlation is very high. Next keep $c = .8$ (correlation = .96) while decreasing τ, the scale of the β's. Figure 7 shows the results. Here, the estimated PE is overly optimistic when τ is 5 or 10 but is reasonably accurate when τ is smaller than 5 or larger than 10. These simulations suggest that (8) is not universally true in the general nonorthogonal case, though it may give a useful approximation. It is a challenging problem to determine how widely these results will continue to hold.

4. THE EVAPORATION EXAMPLE

The accuracy of Formula (8) for the evaporation data has been examined by performing a Monte Carlo study. Note that a naive bootstrap, or simply simulating by adding pseudorandom residuals to the estimated full regression,

This content downloaded on Wed, 6 Feb 2013 16:26:47 PM
All use subject to JSTOR Terms and Conditions

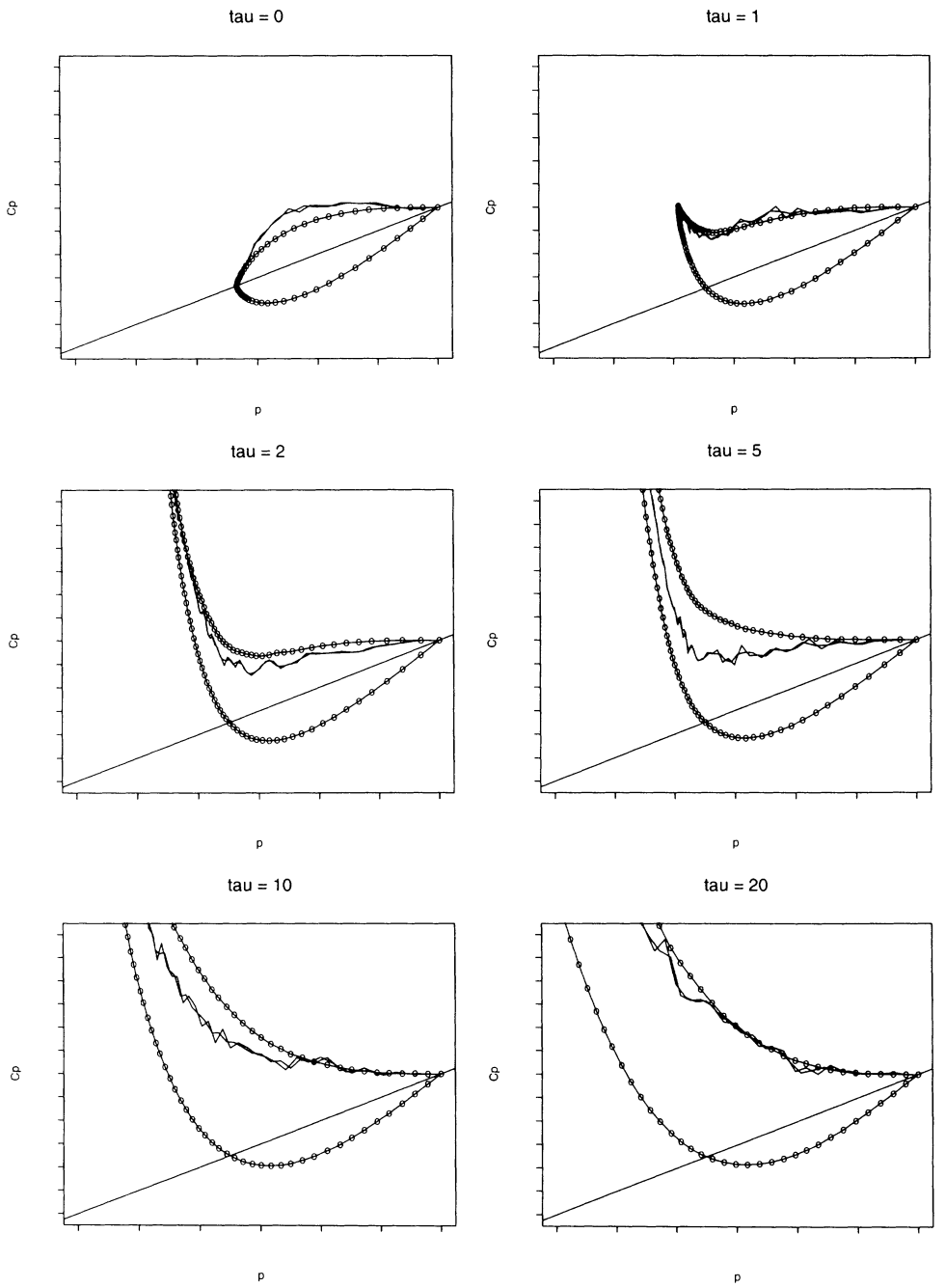

Figure 7. As in Figure 6, With the Within-Pair Correlation Held at .96 and the Scale τ Varying from 0 to 20.

will not work properly because the least squares estimates of the regression coefficients are (probably) larger than the "true" ones. To get a model that approximates what the truth might be, I shrank the coefficients by ridging. First, center and standardize each independent variable so that the diagonal of $A = X^T X$ is the unit matrix I_{10}. Also center the response y. Then fit the full model by least squares, obtaining coefficients $\hat{\beta}$ and an estimate of the residual scale $\hat{\sigma} = 6.508$. Ridged coefficients are obtained by solving $(A + hI_{10})b_h = X^T y$. The ridge parameter h was chosen as follows. Suppose that the vector of "true" regression coefficients has squared length $|\beta|^2$. Then we expect the squared length of b_h to be about $C_1\hat{\sigma}^2 + C_2|\beta|^2$, where $C_1 = \mathrm{tr}(A(A + hI_{10})^{-2})$ and $C_2 = \mathrm{tr}(A^2(A + hI_{10})^{-2})/10$. Thus, h should be chosen to make $C_1/(1 - C_2) = |\beta|^2/\hat{\sigma}^2$. However, $|\beta|^2$ is not known; an unbiased estimate is $|\hat{\beta}|^2 \approx |\beta|^2 + C_3\hat{\sigma}^2$, where $C_3 = \mathrm{tr}(A^{-1})$. Thus we choose h to make $C_1/(1 - C_2) + C_3 = |\hat{\beta}|/\hat{\sigma}^2$. For the evaporation data, $h = .007058$. The values of the standardized least squares and ridged coefficients are given in Table 1. We see that the four largest coefficients have been shrunken considerably but are still so large that they are unlikely to be confused with 0. The smaller coefficients are spread out

This content downloaded on Wed, 6 Feb 2013 16:26:47 PM
All use subject to JSTOR Terms and Conditions

Table 1. Standardized Least Squares and Ridged Regression Coefficients

Regression coefficients	Maxst	Minst	Avst	Maxat	Minat	Avat	Maxh	Minh	Avh	Wing
1s	13.94	.70	−15.37	2.62	1.15	1.99	1.38	7.76	−16.90	1.37
Ridge	8.12	−1.22	−8.67	2.37	.29	4.33	.65	3.36	−12.48	1.97

roughly uniformly. The minimum C_P subset for these data contains variables 1, 3, 6, 8, 9 (and the intercept).

These shrunken coefficients were used to form a "true" response η. Pseudorandom normal variables with scale $\hat{\sigma} = 6.508$ were added; this was repeated until 1,000 cases has been obtained in which the minimizing C_P had $p = 6$ (this took 3,184 cases.) The PE of the minimum C_P subset is

$$\Sigma(\hat{y}_6 - \eta)^2, \tag{15}$$

but to get something directly comparable with C_P we must scale this by $\hat{\sigma}^2$; also the variability of the estimate of PE for the minimum C_P procedure can be reduced by replacing (15) by $11 + (\Sigma(\hat{y}_6 - \eta)^2 - \Sigma(\hat{y}_{11} - \eta)^2)/\hat{\sigma}^2$. The second derivative of the boundary of the C_P plot was estimated by the formula $C_P'' \approx (2C_{p-2} - C_{p-1} - 2C_p - C_{p+1} + 2C_{p+2})/7$. In one case this estimate was negative; in two others it was unreasonably large (greater than 100). For the remaining 997 cases, the estimates in Table 2 were found. We see that Formula (8) is giving at least a rough indication of the inadequacy of the minimum C_P rule. The conclusion is that for the evaporation data set, since PE is estimated to be so large, we should not use minimum C_P (or any other subset-selection rule) for prediction. A similar simulation could be performed for any proposed prediction rule.

5. DISCUSSION

The value of any asymptotic theory is measured by how well it approximates the real (finite) situation. No dramatic change of behavior should occur between the finite case and the asymptotic one. The simulation results encourage belief that this is the case for the present theory; Formula (8), which holds exactly (in the limit) for all $G(\beta)$ in the case of orthogonal regressors and also for uniform β's in the case of correlated regressors with the special structure (14), seems to hold at least approximately as the assumptions are relaxed to allow for (a) general correlated regressors, (b) nonuniform β's, and (c) finite k. I believe that the result (8) [which reduces to (10) for the min C_P rule] is a useful rule of thumb in all cases. It is not clear whether the asymptotics can give a C_P plot with the characteristic "locally uniform" shape of (7) but where the rule (8) is wrong.

Table 2. Means and Standard Deviations of the Minimum C_P, Formula (8), and the Estimated PE, for the Simulated Evaporation Data

	Mean	Standard deviation
C_P	3.457	1.165
(8)	10.683	6.508
PE	11.471	2.314
PE–(8)	.788	7.021

A referee has suggested that the distribution of the predictors may be relevant; for example, he expects (8) to fail when the conditional expectations are nonlinear. I do not see why this should be so.

If the asymptotic theory is a reasonable approximation to a real situation, the implication for prediction is clear. Subset–least squares should not be used, at least whenever the C_P plot has a broad minimum like Figure 1. Fortunately, several alternatives are available. In classical ridge regression (Hoerl and Kennard 1970), we choose a parameter h and set $\hat{y} = L_h y$, where $L_h = X(X^T X + hI)^{-1} X^T$. For any fixed value of the parameter h, this is a linear procedure. Mallows (1973) defined statistic C_L that estimates the PE of an arbitrary linear estimator; it was pointed out that even when the ridge parameter r is chosen to minimize C_{L_r}, this minimized value does give a reasonable estimate of the corresponding PE. It is not clear how to provide similar estimates for more general nonlinear procedures.

The ridge procedure shrinks all of the coefficients toward 0. Perhaps only the "small" ones should be shrunk and the large ones left alone. A heuristic proposal along these lines was made by Mallows (1973). Mitchell and Beauchamp (1988) provided a Bayesian analysis that achieved this result; it assumed a prior of "spike + slab" form, depending on a parameter γ, which is the ratio of the height of the spike to the height of the slab. They described a battery of graphical procedures that help one understand the sensitivity of the Bayes procedure to the choice of γ. It would appear reasonable to estimate γ from the data and to use the corresponding Bayes estimate (which is a weighted average of the subset–least squares estimates) for prediction. It is not clear (to me) whether the Bayes formula that has been derived for the PE of the fixed-γ procedure will provide a useful estimate of the PE of the estimated-γ procedure.

The Gibbs sampling algorithm of George and McCulloch (1993) provided a convenient methodology for these Bayesian computations when the number of variables is large. A simple approximation to the Bayesian procedure is to average the subset predictions as follows:

$$\hat{\beta} = \sum_P \hat{\beta}_P e^{-1/2C_P} |X_P^T X_P|^{-1/2}.$$

This content downloaded on Wed, 6 Feb 2013 16:26:47 PM
All use subject to JSTOR Terms and Conditions

Miller (1990) discussed several nonlinear proposals. Recently Breiman (1995) suggested a *nonnegative garotte* in which the ordinary least squares regression estimates are shrunk by factors whose sum is constrained. Tibshirani (1994) suggested a procedure named *lasso* in which the sum of the absolute values of regression coefficients themselves is constrained. In both cases, the value of the constraint is to be chosen by a cross-validation computation.

ACKNOWLEDGMENTS

The comments of two referees and an associate editor were very helpful in revising an earlier version.

APPENDIX A: DERIVATION OF (8) IN THE CASE OF ORTHOGONAL REGRESSORS

From assumption (5), we can derive the asymptotic configuration of the b's:

$$\frac{1}{k'}\sum[b_j < t] \to \int \Phi(t-\beta)\,dG(\beta),$$

where Φ is the standard Gaussian cumulative distribution. (Write ϕ for the corresponding density.) For any $t > 0$, the subset $P(t) = \{j: |b_j| > t\}$ has

$$|P(t)| \approx k - k' \int dG(\beta) \int_{-t}^{t} \phi(b-\beta)\,db$$

and minimizes C_P among subsets of this size; moreover, since the residual sum of squares for this fit is $\sum_{Q(t)} b_j^2$, using (3) we have

$$\begin{aligned} C_{P(t)} &= k + \sum_{j\in Q(t)} \left(b_j^2 - 2\right) \\ &\approx k + k' \int dG(\beta) \int_{-t}^{t} (b^2-2)\phi(b-\beta)\,db. \end{aligned} \quad \text{(A.1)}$$

In the "asymptotically uniform" case—that is, with G given by (6) with local density $\lambda = k'/2\tau$—we find $|P(t)| = k - q \approx k - 2t\lambda$ and (7) follows easily.

Suppose that I use least squares to estimate the coefficients $b_{P(t)}$ and hence predict the responses at n new observations, where the design of the new observations is the same as that in the data in hand. Thus, the predictions are $\hat{y} = X_{P(t)} b_{P(t)}$. The prediction error (ignoring the part $|\eta - X\beta|^2$) is

$$\mathrm{PE}_{P(t)} = \sum_{Q(t)} \beta^2 + \sum_{P(t)} (\beta - b)^2,$$

which has expectation

$$\begin{aligned} E(\mathrm{PE}_{P(t)}) &\approx k + E\sum_{Q(t)} (2\beta b - b^2) \\ &\approx k + k' \int dG(\beta) \int_{-t}^{t} (2\beta b - b^2)\phi(b-\beta)\,db. \end{aligned}$$

In the asymptotically uniform case, this is $k + \frac{2}{3}\lambda t^3$. For general G, simple calculus, treating p as a continuous variable, gives [using (7)]

$$\begin{aligned} E(\mathrm{PE}_{P(t)}) - C_{P(t)} &\approx 2tk' \int \{\phi(t-\beta) + \phi(t+\beta)\}\,dG(\beta) \\ &= -2t\frac{dp}{dt}, \end{aligned} \quad \text{(A.2)}$$

whereas from (A.1) we have $dC_{P(t)}/dp \approx 2 - t^2 = C'_{P(t)}$, say, so that $d^2C_{P(t)}/dp^2 \approx -2t(dt/dp) = C''_{P(t)}$, say. Formula (8) follows from (A.2).

An exactly parallel development goes through if we assume that the independent variables are not standardized, so that $y_i = x_i\beta_i + e_i$, where the e's are iid Gaussian as previously, and now the limiting configuration of the β's may depend on x; that is, I now assume that

$$\frac{1}{k}\sum[x_i < \xi] \to F(\xi)$$

and

$$\frac{1}{k}\sum[x_i < \xi \,\&\, \beta_i < t] \to \int_0^{\xi} dF(x)G(t \mid x).$$

In this case we have

$$\begin{aligned} \frac{1}{k}|Q(t)| &\to \int dF(x) \int dG(\beta \mid x) \int_{-t}^{t} \phi(y - x\beta)\,dy, \\ \frac{1}{k}C_{P(t)} &\to 1 + \int dF(x) \int dG(\beta \mid x) \\ &\qquad \times \int_{-t}^{t} (y^2 - 2)\phi(y - x\beta)\,dy, \end{aligned}$$

and

$$\begin{aligned} \frac{1}{k}E(\mathrm{PE}(t)) &\to 1 + \int dF(x) \int dG(\beta \mid x) \\ &\qquad \times \int_{-t}^{t} (2x\beta y - y^2)\phi(y - x\beta)\,dy, \end{aligned}$$

and we can check that the relation (8) continues to hold.

APPENDIX B: DERIVATION OF (8) IN THE NONORTHOGONAL CASE, LOCALLY UNIFORM β'S

Assume that design has the structure (13). We may assume that (by linear operations on y) X has been reduced to a square nonsingular matrix. Then we may write $y^T = (y_1^T, \ldots, y_m^T)$, where each subvector $y_j = X\beta_j + e_j$ is $k \times 1$, X is now $k \times k$, and β_j is $k \times 1$, $j = 1, 2, \ldots, m$. The least squares estimate of β_j is $b_j = X^{-1}y_j$. Let X_i denote the ith column of X.

Assume that, as $m \to \infty$, the empirical distribution of the β_j's approaches a Gaussian distribution with mean 0 and covariance $\tau^2 V$, where τ is large (relative to unity). Write the (k-dimensional) differential of this measure as $d\Phi(\cdot; \tau^2 V)$. Given β_j, y_j is standard Gaussian with mean $X\beta_j$; that is, its density is $\phi(y - X\beta)$. We have $b_j = X^{-1}y_j$ so that the empirical distribution of the b_j's approaches $\Phi(\cdot; \tau^2 V + (X^T X)^{-1})$.

This content downloaded on Wed, 6 Feb 2013 16:26:47 PM
All use subject to JSTOR Terms and Conditions

For any subset P of $K = \{1, \ldots, k\}$, let X_P be the corresponding submatrix of X, and let $H_P = X_P \times (X_P^T X_P)^{-1} X_P^T$ be the corresponding projection matrix. Write $\overline{H}_P = I_k - H_P$. Now fix a value of t, and consider the rule that minimizes

$$C_P - mk + (t^2 - 2)p = \sum_i \mathrm{SS}_{i,P_i} - t^2 q_i$$
$$= \sum_i y_i \overline{H}_{P_i} y_i - t^2 q_i. \quad \text{(B.1)}$$

This rule finds the "lower boundary" subset, where the tangent to the cloud of (p, C_P) points has slope $2 - t^2$. To minimize the quantity in (B.1), we must minimize the contribution from each block separately. For each subset P, let $R_P(t)$ be the region in R^k within which the subset P is selected, for our chosen value of h. Thus, $R_P(t) = \{y\colon \text{for all } P',\ y^T \overline{H}_P y - t^2|Q| \le y^T \overline{H}_{P'} y - t^2|Q'|\}$. As $m \to \infty$, the fraction of blocks for which a particular subset P is selected is thus

$$\int d\Phi(\beta; \tau^2 V) \int_{R_P(t)} d\Phi(y - X\beta).$$

Hence, as $m \to \infty$,

$$q \approx m \int d\Phi(\beta; \tau^2 V) \sum_P q \int_{R_P(t)} d\Phi(y - X\beta)$$
$$= m \sum_P q \int_{R_P(t)} d\Phi(y; I + \tau^2 X V X^T). \quad \text{(B.2)}$$

For subsets on the lower boundary of the C_P plot,

$$C_P - mk + (t^2 - 2)p$$
$$\approx m \int d\Phi(\beta; \tau^2 V) \sum_P \int_{R_P(t)} d\Phi(y - X\beta) \times \left(y^T \overline{H}_P y - t^2 q\right)$$
$$= m \sum_P \int_{R_P(t)} d\Phi(y; I + \tau^2 X V X^T)\left(y^T \overline{H}_P y - t^2 q\right), \quad \text{(B.3)}$$

and the corresponding excepted predictive error is given by

$$E(\mathrm{PE}) - mk \approx m \int d\Phi(\beta; \tau^2 V) \sum_P \int_{R_P(t)} d\Phi(y - X\beta) \times \left(|X\beta - X_P \hat{\beta}_P|^2 - |X\beta - y|^2\right). \quad \text{(B.4)}$$

Each of the expressions (B.2), (B.3), and (B.4) is of order m/τ. The crucial step in the argument is the observation that in each of these expressions we can ignore the term with $|P| = k$ because integrands vanish identically; we can also ignore all terms with $|P| \le k - 2$ because these contribute only quantities of order m/τ^2. The only terms that contribute nonnegligibly are those with $|P| = k - 1$. For each j, write $P_j = K - \{j\}$. Then, in each case, the term involving P_j can be approximated by replacing the region R_{P_j} by

$$R_j^* = \{y\colon y^T \overline{H}_P y - t^2 < 0\}$$
$$= \{y\colon -t < \xi_j^T y < t\},$$

where ξ_j is a unit vector satisfying $\xi_j \xi_j^T = \overline{H}_{P_j}$. Set $w_j = \xi_j^T y$, and define v_j by $v_j^2 = \mathrm{var}(w_j) = 1 + \tau^2 (\xi_j^T X_j)^2 V_{jj}$. Note that $(\xi_j^T X_j)^2 = X_j^T \overline{H}_{P_j} X_j = (1 - r_j^2)|X_j|^2$, where r_j is the multiple correlation between X_j and the other X's. Thus,

$$q \approx m \sum_j \int_{-t}^{t} \phi\left(\frac{w_j}{v_j}\right) \frac{dw_j}{v_j} \approx m \sum_j \frac{2h}{\sqrt{2\pi} v_j} \approx 2h\lambda,$$

where

$$\lambda = \frac{m}{\sqrt{2\pi}\tau} \sum_{j=1}^{k} \frac{1}{v_j}.$$

Similarly,

$$C_P - mk + (t^2 - 2)p$$
$$\approx m \sum_j \int_{-t}^{t} (w_j^2 - h^2) \phi\left(\frac{w_j}{v_j}\right) \frac{dw_j}{v_j}$$
$$\approx \left(\frac{2}{3} t^3 - 4t\right) \lambda$$

so that the shape of the lower boundary of the C_P plot is the same as in the orthogonal case.

Similarly we can approximate the predictive error. First, the integrand in (A.6) simplifies to $2\beta^T X^T \overline{H}_{P_j} y - y^T \overline{H}_{P_j} y$. Given y, β is Gaussian with mean $W_\tau y$ and covariance W_τ, where $W_\tau = \tau^2 X V X^T (I + \tau^2 X V X^T)^{-1}$. Note that $W_\tau = I + O(\tau^{-2})$. Hence (B.4) becomes approximately

$$m \sum_j \int_{R_j(t)^*} y^T (2W_\tau - I) \overline{H}_{P_j} y \, dG(y; I + \tau^2 X V X^T)$$
$$\approx m \sum_j \int_{-t}^{t} w_j^2 \phi\left(\frac{w_j}{v_j}\right) \frac{dw_j}{v_j} \approx \frac{2}{3} h^3 \lambda$$

as in the orthogonal case. Result (8) applies.

[*Received April 1994. Revised February 1995.*]

REFERENCES

Becker, R. A., Chambers, J. M., and Wilks, A. R. (1988), *The New S Language*, Pacific Grove, CA: Wadsworth & Brooks Cole.

Breiman, L. (1995), "Better Subset Selection Using the Nonnegative Garotte," *Technometrics*, 37, 373–384.

Breiman, L., and Freedman, D. (1983), "How Many Variables Should Be Entered in a Regression Equation?" *Journal of the American Statistical Association*, 78, 131–136.

Freund, R. J. (1979), "Multicollinearity, etc., Some 'New' Examples," in *Proceedings of the Statistical Computing Section, American Statistical Association*, pp. 111–112.

George, E. I., and McCulloch, R. E. (1993), "Variable Selection via Gibbs Sampling," *Journal of the American Statistical Association*, 88, 881–889.

Hoerl, A. E., and Kennard, R. W. (1973), "Ridge Regression: Biased Estimation for Non-orthogonal Problems," *Technometrics*, 12, 55–67.

This content downloaded on Wed, 6 Feb 2013 16:26:47 PM
All use subject to JSTOR Terms and Conditions

Mallows, C. L. (1973), "Some Comments on *Cp*," *Technometrics*, 15, 661–675.

Miller, A. J. (1990), *Subset Selection in Regression*, New York: Chapman and Hall.

Mitchell, T. J., and Beauchamp, J. J. (1988), "Bayesian Variable Selection in Linear Regression," *Journal of the American Statistical Association*, 83, 1023–1032.

Schwartz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 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," *Journal of the Royal Statistical Society*, Ser. B, 39, 44–47.

Tibshirani, R. (1994), "Regression Shrinkage and Selection via the Lasso," technical report, University of Toronto, Dept. of Biostatistics.

This content downloaded on Wed, 6 Feb 2013 16:26:47 PM
All use subject to JSTOR Terms and Conditions