COLLINEARITY AND STABILITY IN THE ESTIMATION OF RAINFALL–RUNOFF MODEL PARAMETERS

PETER K. KITANIDIS and RAFAEL L. BRAS *

Department of Civil Engineering, University of Iowa, Iowa City, IA 52240 (U.S.A.)
Department of Civil Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139 (U.S.A.)

(Received August 10, 1978; revised and accepted December 28, 1978)

ABSTRACT


The common use of statistical rainfall–runoff models of the regressive type lead to the problem of handling the unavoidable correlation existing between the explanatory variables of these models. The above collinearity problem is treated using ridge regression techniques. This powerful regression method is extensively presented and discussed. Examples show that collinearity has significant effects on values of rainfall–runoff model parameters and that ridge regression effectively deals with the problem, leading to well behaved model coefficients. The important issue of parameter stability is also addressed using both ridge regression and Kalman filtering as estimation procedures. It is shown that considerable data are required before parameter stability is achieved casting serious doubts on the predictive capability of models developed with usually limited historical data. Although a powerful methodology, Kalman filtering exhibited disturbing instability and lack of convergence unless special algorithms of high numerical accuracy are used.

INTRODUCTION

The regression exercises reported in the following paper attempted to address known problems of hydrologic models of the type suggested by Todini et al. (1976). In particular, we studied:

(1) Collinearity problems resulting from the unavoidable correlation between explanatory variables in statistical rainfall–runoff models. This issue was studied using recent ridge regression concepts.

(2) Stability of parameters in regression models. The question concerns the length of data required to obtain stable regression coefficients which are logical and trustworthy. Parameter stability was studied with ridge regression and Kalman filter estimation approaches.

*To whom correspondence should be addressed.
(3) Stability of Kalman filter. Even though a relatively stable version of the Kalman filter, known as the Potter algorithm, was used, it would not converge unless further conditioning of the data was used. It was found that standardization of all independent variables to the same order of magnitude caused convergence and satisfactory accuracy of the Kalman filter results.

It was concluded that:

1. Collinearity has significant effects on values of regression parameters in rainfall—runoff models.
2. Ridge regression yields results, under the parameter stability criteria, leading to reasonably well behaved regression coefficients.
3. Considerable data are needed before parameter stability is achieved. Kalman filter estimation techniques showed that after 4 years of daily streamflow data regression coefficients were still varying to a noticeable extent.

The following sections discuss in detail the work that lead to the above conclusions.

COLLINEARITY

Collinearity is one of the problems arising in the use of regression models for the prediction of one-day-ahead daily river discharges. For example, Kashyap and Rao (1973) and others have used among their explanatory variables the flows measured one and two days before the day of the predicted flow. These two variables are highly correlated. For large rivers with relatively long response time, the correlation coefficient may be as high as 0.95 while for smaller rivers it is of the order of 0.85. One of the most promising methods in coping with the problems of collinearity is ridge regression. According to the results of a recent simulation study (Dempster et al., 1977):

"The substantive conclusions from this study are indications of possible drastic improvements over (ordinary) L.S., especially through the technique of ridge regression, and especially when a high degree of correlation exists among the independent variables."

From a frequentist's standpoint, it has long been recognized that mean squared error properties do not necessarily follow from the minimum variance unbiasedness properties of ordinary least squares. Allowing for bias may improve the total estimation error, since in certain regions of the parameter space, the loss from increasing the squared bias can be over-compensated by reducing variance. Existence theorems have been presented by Hoerl and Kennard (1970a,b), and Theobald (1974).

From a Bayesian point of view, ridge regression can be considered as a special case of Bayesian regression. Although lacking in generality, it has the merit that it depends more on the given data to provide us with information, it gives us a better feeling on what the stability of the estimates is, and requires less prior information about the parameters than ordinary Bayesian regression.
ORDINARY LEAST SQUARES AND COLLINEARITY

Suppose we have the standard linear regression problem:
\[ Y = X\beta + \epsilon \]  \hspace{1cm} (1)

where \( X \) is \( (n \times m) \) and of rank \( m \), \( \beta \) is \( (m \times 1) \) and unknown, \( E(\epsilon) = 0 \) and \( E(\epsilon \epsilon^T) = \sigma^2 I_n \).

The usual estimation procedure for the unknown parameters \( \beta \) uses an estimate \( \hat{\beta} \) which is a linear function of \( Y \) and satisfies the following conditions: (1) the estimates are unbiased; and (2) have minimum variance.

Hoerl and Kennard argue that this estimation procedure is good only if \( X^TX \) is close to a diagonal matrix. Otherwise, the least-squares estimates are "sensitive to a number of errors".

The unbiased minimum variance estimate of the parameters is given:
\[ b_{LS} = (X^TX)^{-1}X^TY \]  \hspace{1cm} (2)

with estimation error covariance:
\[ \text{Var}(b_{LS}) = \sigma^2(X^TX)^{-1} \]  \hspace{1cm} (3)

To examine the properties of this estimate in a convenient way, Hoerl and Kennard (1970a) used the Euclidean distance \( L_1 \) of \( b_{LS} \) from \( \beta \):
\[ L_1^2 = (b_{LS} - \beta)^T(b_{LS} - \beta) \]  \hspace{1cm} (4)

Since \( L_1^2 \) is not known, by taking its expected value:
\[ E(L_1^2) = E[(b_{LS} - \beta)^T(b_{LS} - \beta)] = \text{trace}[E(b_{LS} - \beta)(b_{LS} - \beta)^T] \]
\[ = \sigma^2 \sum_{i=1}^{m} 1/\lambda_i \]  \hspace{1cm} (5)

where \( \lambda_i \)'s are the eigenvalues of \( (X^TX) \). However, \( X^TX \) is positive definite, so that all its eigenvalues are positive, summing up to the trace of \( (X^TX) \). If \( X \) has been normalized so that \( X^TX \) is a correlation matrix, and if \( X^TX \) is close to a diagonal matrix, then \( \lambda_i = 1 \), and:
\[ E(L_1^2) \approx \sigma^2 m \]  \hspace{1cm} (6)

However, as \( X^TX \) moves from a diagonal unit matrix to an ill-conditioned one, the ratio \( \lambda_{\max}/\lambda_{\min} \) (its square root is called the condition number), where:
\[ \lambda_{\min} = \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_m = \lambda_{\max} \]  \hspace{1cm} (7)

increases dramatically. Since all \( \lambda_i \)'s must sum up to the trace of \( X^TX \), \( \lambda_{\min} \) can become very small so that \( E(L_1^2) \) (eq. 5) can become very large. That explains why, in the presence of collinearity, a much too long parameter vector is estimated, since the expected value of the Euclidean length of the param-

The vector is:

\[ E(b_{LS}^T b_{LS}) = \beta^T \beta + \sigma^2 \sum \frac{1}{\lambda_i} \]  

(8)

RIDGE REGRESSION

Before we continue our analysis, it is convenient to transform the data so that \( X^T X \) will be a correlation matrix. As noted by Mallows (1973), this can be achieved by centering \( X \) and \( Y \) and by properly standardizing the columns of \( X \). This facilitates the analysis and at the same time serves in introducing the right scaling between the coefficients.

For hydrologic modeling, we are generally interested in regressions through the origin, i.e., with zero constant term (intercept). It is easy to show that under the standard linear regression assumptions \( X^T X \) is a positive definite matrix which by the necessary data transformations becomes a correlation matrix. Although \( X^T X \) is not the correlation matrix of the independent variables, it is the one that appears in the calculations and, consequently, the one we are interested in.

Hoerl first suggested in 1962 that:

"in order to control the inflation and general instability associated with the least squares estimates"

(Hoerl and Kennard, 1970a) one can use:

\[ b_k = (X^T X + kI)^{-1} X^T Y, \quad k \geq 0 \]  

(9)

Thus, we get a parametric family of linear estimators, called ridge estimators, a special case of which is the unbiased minimum variance estimator (ordinary least squares) obtained for \( k = 0 \). These estimators can be given a nice Bayesian interpretation which will become clear in a following section. For the time being, let us examine their justification from a frequentist’s standpoint.

From a practical point of view, let us see how we reduce the effects of collinearity by taking \( k > 0 \).

Since \( X^T X \) is positive definite, it can be decomposed through the use of Householder’s orthogonal matrices:

\[ X^T X = VEV^T \]  

(10)

where \( V \) is orthonormal \( (V \cdot V^T = 1) \) and \( E \) is the diagonal matrix of the eigenvalues \( \lambda_i \) of \( X^T X \). Then:

\[ X^T X + kI = VEV^T + kVV^T = V[E + kI]V^T \]  

(11)

Since \( [E + kI] \) is still diagonal and \( V \) is orthonormal, the inverse can be easily found:

\[ [X^T X + kI]^{-1} = V[E + kI]^{-1}V^T \]  

(12)

where \( [E + kI]^{-1} \) is the diagonal matrix with elements \( 1/(\lambda_i + k) \).
From eqs. 2 and 9, we get the relationship between the ridge and the OLS (ordinary least-squares) estimate:

\[ b_k = (X^T X + kI)^{-1} X^T X b_{LS} \]  

which can conveniently be written as:

\[ b_k = V[\frac{\lambda_i}{(\lambda_i + k)}] V^T b_{LS} \]

where \( [\frac{\lambda_i}{(\lambda_i + k)}] \) is a diagonal matrix with elements \( \frac{\lambda_i}{(\lambda_i + k)} \). Thus, it becomes obvious that the vector of the ridge regression parameter estimates is shorter than the vector of the least-squares one.

In this way, the overinflation of the parameter estimates in OLS is effectively handled. It turns out, however, that theoretically the ridge estimates of the parameters can also be better than the OLS ones. Hoerl and Kennard (1970a) proved the following existence theorem:

**Theorem.** Let us define the expected value of the square of the Euclidean distance of the vector of the ridge parameter estimates from the vector of the true parameters:

\[ E\{L^2(k)\} = E[(b_k - \beta)^T (b_k - \beta)] \]

Then, there always exists a \( k > 0 \), such that:

\[ E\{L^2(k)\} < E\{L^2(0)\} = \sigma^2 \sum_{i=1}^{m} \frac{1}{\lambda_i} \]

This theorem provided a first theoretical justification of ridge regression. Since then, Hoerl and Kennard’s argument has been criticized for the use of the Euclidean distance (see, e.g., Nelder, 1972). In reality, we should look at second moment matrices in order to provide an improved sampling-theoretic justification for the ridge regression. For that, we follow the analysis of Theobald (1974).

We define the second-moment matrices corresponding to the estimators \( \tilde{\theta}_1 \) and \( \tilde{\theta}_2 \) of \( \theta \):

\[ M_j = E[(\tilde{\theta}_j - \theta)(\tilde{\theta}_j - \theta)^T], \quad j = 1, 2 \]

We will consider estimator \( \tilde{\theta}_1 \) superior over estimator \( \tilde{\theta}_2 \) if and only if \( M_1 \) is “smaller” than \( M_2 \) in the sense:

\[ M_2 - M_1 = \text{nonnegative definite} \]

Let us further define two scalar functions of an arbitrary nonnegative definite matrix \( B \):

\[ m_j(B) = E[(\tilde{\theta}_j - \theta)^T B(\theta_j - \theta)], \quad j = 1, 2 \]

The following theorem can be proven (see Theobald, 1970):
Theorem. The following two conditions are equivalent:

1. \( M_2 - M_1 \) is nonnegative definite
2. \( m_2(B) - m_1(B) \geq 0 \), for all n.n.d. \( B \)

Thus, superiority of \( \hat{\theta}_1 \) over \( \hat{\theta}_2 \) is not that:

\[
E[(\hat{\theta}_1 - \theta)^T(\hat{\theta}_1 - \theta)] \leq E[(\hat{\theta}_2 - \theta)^T(\hat{\theta}_2 - \theta)]
\]

as implied by Hoerl and Kennard (1970a) but rather that:

\[
E[(\hat{\theta}_1 - \theta)^TB(\hat{\theta}_1 - \theta)] \leq E[(\hat{\theta}_2 - \theta)^TB(\hat{\theta}_2 - \theta)]
\]

for every n.n.d. \( B \).

With respect to the ridge estimator of the parameters:

\[
b(k) = (X^TX + kI)^{-1}X^TY, \quad k \geq 0
\]

the second moment (estimation error covariance matrix) will be composed of two terms, the one corresponding to the bias:

\[
c(k) = -k(X^TX + kI)^{-1}\beta
\]

and the other corresponding to the dispersion around the mean:

\[
D(k) = \sigma^2 X^TX(X^TX + kI)^{-2}
\]

Thus the second-order moment is:

\[
M(k) = D(k) + c(k)c^T(k)
\]

For this second moment, Theobald proved that: there exists a \( K > 0 \), such that \( M(0) - M(k) \) is positive definite whenever \( 0 < k < K \).

This theorem provides a solid theoretical justification for ridge regression. However, it is only an existence theorem and does not give us any information about what values \( k \) takes.

RIDGE TRACE AND MALLOWS' \( C_k \)

Ridge regression can be used to provide valuable information about the stability of the parameter estimates in the presence of collinearity. The most commonly used tool is the ridge trace, that is the plot of the estimates \( b_i(k) \) as a function of \( k \). Hoerl and Kennard (1970b) proposed the selection of that \( k \) for which the \( b_i(k) \) has started to stabilize.

Another tool that proved useful in the present analysis is a tradeoff curve between fit (\( RSS_k \)) and the square of the length of the parameters vector (\( b^Tb \)). By using this, the decision maker can decide that he wants to trade, for example, 4\% of the fit as expressed by the RSS for a 50\% reduction of \( b^Tb \).

These are useful sensitivity tools but do not lead to a single value of \( k \). A methodology for selecting a single value of \( k \) as optimum has been proposed by Mallows (1973). It consists essentially of optimizing a statistic named \( C_k \), which is very similar to the \( C_p \) statistic used in the all possible regressions approach for variable selection.
To define $C_k$, Mallows uses as measure of the predictive quality of the ridge parameter estimate the scaled summed mean-square error,

$$J_k = \sigma^{-2} (Xb_k - X\beta)^T(Xb_k - X\beta)$$

(21)

Mallows showed that the expected value of $J_k$ is the sum of a bias and a covariance term. $C_k$ is a statistic, which is a function of the data, with the same expected value as $J_k$. Consequently, by finding the $k$ which minimizes $C_k$, we expect to minimize in some average sense the mean square error $J_k$.

Optimization of $C_k$ is straightforward. However, it is not clear what is the variance of $C_k$ so that it is not perfectly clear how meaningful trying to optimize this statistic would be. Additionally, on theoretical grounds, one may object to the selection of $J_k$ as the criterion. In the previous section, we discussed the criterion for comparing estimators (comparison between second-moment matrices rather than between Euclidean distances). One can only consider $J_k$ as a "reasonable" criterion and the resulting solution as a "reasonably good" one.

Mallows notes that for the example presented by Hoerl and Kennard (1970b), the $C_k$ criterion suggests a value approximately 10 times smaller than the value suggested by their stabilization criterion. In all cases examined in this work, optimization of Mallows' $C_k$ gave values approximately 10–1000 times smaller than what one would expect when using a stabilization criterion. In some cases, the $C_k$ method suggested values that were so small that it would make the ridge regression solution practically unrecognizable from the OLS solution, even though there was some collinearity.

RIDGE AND BAYESIAN REGRESSION THROUGH DATA AUGMENTATION

The purpose of this section is to illustrate the direct relation between Bayesian and ridge regression and at the same time provide a useful way of applying them through a usual OLS program. The discussion will also illuminate certain theoretical aspects.

Suppose that we have a priori unbiased estimate of $\beta$, $\tilde{\beta}$, and an associated error covariance matrix, $\tilde{P}$ (or information matrix $\Lambda = \tilde{P}^{-1}$). We will now show how we can in fact incorporate this information in the form of additional data points.

We can in general factor the positive definite (or even more generally positive semi-definite) symmetric information matrix into:

$$\tilde{\Lambda} = \tilde{R}^T\tilde{R}$$

(22)

We can, for example, use Householder decomposition:

$$\tilde{\Lambda} = V E V^T = (VE^{1/2})(E^{1/2}V^T) = (E^{1/2}V^T)T(E^{1/2}V^T)$$

(23)

The loglikelihood functional $J(b)$, modified to account for prior information, is:

$$J(b) = (\tilde{b} - b)\tilde{\Lambda}(\tilde{b} - b) + (y - Xb)^T(y - Xb)\sigma^{-2}$$

(24)
We can introduce vector \( \tilde{y} \):
\[
\tilde{y} = \tilde{R} \tilde{b}
\]
and \( J(b) \) can be rewritten as:
\[
J(b) = (\tilde{y} - \tilde{R}b)^T(\tilde{y} - \tilde{R}b) + (y - Xb)^T(y - Xb) \cdot \sigma^{-2}
\]
(26)

This is equivalent to having added to our original data a number of observations equal to the number of the parameters to be estimated. Therefore, we can interpret the least-squares performance functional as corresponding to the data set:
\[
\begin{bmatrix}
\tilde{y} \\
y/\sigma
\end{bmatrix} = \begin{bmatrix}
\tilde{R} \\
X/\sigma
\end{bmatrix} \beta + \begin{bmatrix}
\tilde{e} \\
\epsilon/\sigma
\end{bmatrix}
\]
(27)

where division by \( \sigma \) was used in order to normalize all residuals to unit variance, since \( \tilde{e} \sim N(0, 1) \), \( \epsilon \sim N(0, \sigma^2) \). Direct solution with the well-known ordinary least-squares formulas:
\[
b = \left[ \tilde{R}^T, \frac{X^T}{\sigma} \right] \left[ \tilde{R} \right]^{-1} \left[ \tilde{R}^T, \frac{X^T}{\sigma} \right] \begin{bmatrix} \tilde{y} \end{bmatrix} = \left[ \frac{X^T X}{\sigma^2} + \tilde{R}^T \tilde{R} \right]^{-1} \left[ \tilde{R}^T, \frac{X^T}{\sigma} \right] \begin{bmatrix} \tilde{y} \end{bmatrix}
\]
\[
= \left[ \frac{X^T X}{\sigma^2} + \Lambda \right]^{-1} \Lambda b + \left[ \frac{X^T X}{\sigma^2} + \Lambda \right]^{-1} X^T \tilde{y} \quad (28)
\]

In the Bayesian model and under the assumptions that the prior estimate is unbiased and uncorrelated from the given data set, this estimator is unbiased and of minimum variance. The covariance matrix is:
\[
P = \left[ \Lambda + \frac{X^T X}{\sigma^2} \right]^{-1}
\]
(29)

For the case of \( \tilde{b} = 0 \), \( \Lambda = \tau^{-2} \cdot I \) (where \( \tau^{-2} \), as we will see later, can be interpreted as the inverse of a prior variance) and \( X^T X = \text{correlation matrix} \), we have a ridge estimator. In this case, \( \tilde{R} = \tilde{R}^T = \tau^{-1} \cdot I \) and \( \tilde{y} = 0 \). Then the artificial additional points are:
\[
0 = \tau^{-1} \beta_i + \tilde{e}_i, \quad i = 1, m
\]
(30)
or
\[
\beta_i = -\tau \cdot \tilde{e}_i
\]
(31)

where \( \tilde{e}_i \) is an \( N(0, 1) \) random variable.

Consequently, ridge regression essentially represents prior information on the form of each parameter being independent from any other and Gaussian with zero mean and variance \( \tau^2 \). For \( \tau^2 = \infty \), we have noninformative prior distribution of the parameters. However, it may be argued that in most cases
it is much more likely that a parameter is in the interval \([-M/2, M/2]\) rather than in the interval \([M/2, 3M/2]\) or \([-3M/2, -M/2]\), for \(M\) a sufficiently large number. Yet it is known that in the presence of collinearity, inverting the almost singular matrix \((X^TX)\) will tend to produce estimates of the parameters that are arbitrarily large in absolute value. Ridge regression appears as a way of conditioning the matrix \(X^TX\) by using in the estimation the type of prior information we have described.

In practice, instead of using the prior variance \(\tau^2\) and the, in general, not known a priori, \(\sigma^2\), we use the ratio:

\[
k = \frac{\sigma^2}{\tau^2}
\]

Then:

\[
b = (X^TX + kI)^{-1}X^T\gamma \quad \text{and} \quad P = (X^TX + kI)^{-1}\sigma^2
\]

And since \(\gamma = X\beta + \epsilon\) is usually not scaled so that \(\epsilon\) has variance one, the equivalent data set is:

\[
\begin{bmatrix}
\tilde{y} \\
\tilde{\gamma}
\end{bmatrix} = \begin{bmatrix}
\tilde{R} \\
X
\end{bmatrix} \beta + \begin{bmatrix}
\sigma \\
\tilde{\epsilon}
\end{bmatrix}
\]

with

\[\epsilon \sim N(0, \sigma^2) \quad \text{and} \quad \tilde{\epsilon} \sim N(0, 1)\]

Then the artificial data that are simply added to the original data are:

\[0 = k^{\frac{1}{2}}\beta_i, \quad i = 1, \ldots, m\]

ESTIMATION OF PARAMETERS OF MODELS FOR DAILY RIVER FLOW PREDICTION THROUGH RIDGE REGRESSION

We will now analyze the daily flow data for Cohocton River, measured near Campbell, N.Y., and we will try to postulate a model useful for predictive purposes. Approximately 8 years of data (2828 days) of mean daily flows and mean areal precipitation of the river catchment were available. Since, however, this is more data than is often available, we will also analyze separately the data of the first year of measurements, 1964 (365 days). Analysis was performed using an efficient algorithm developed to perform ordinary least-squares and ridge regression.

In an exploratory fashion, we will postulate a model of the form:

\[
Q(t) = \beta_1 Q(t-1) + \beta_2 Q(t-2) + \beta_3 i(t) + \beta_4 i(t-1) + \beta_5 i(t-2) + \beta_6 Q^2(t-1) \\
+ \beta_7 Q(t-1)i(t-1) + \beta_8 Q(t-2)i(t-2)
\]

where \(Q(t) = \text{mean daily flow (ft.}^3\text{s}^{-1}, i(t) = \text{mean areal rainfall (mm)}\).

Let us first fit this model to the 365 first cases, corresponding to calendar
year 1964. By looking at the correlation matrix, it becomes obvious that we have significant interfactor correlations. \( Q(t-1) \) and \( Q(t-2) \) have a correlation coefficient of 0.763, \( Q(t-1) \) and \( Q^2(t-1) \) a coefficient of 0.8463, \( Q(t-1) \) and \( Q(t-1) \cdot i(t-1) \) a coefficient of 0.63. In general, the correlation matrix is pretty "dense" and the condition number, as a measure of near collinearity for the OLS case, is 11.01, which is not high. Notice, however, that without data normalization and singular value decomposition the condition number would be orders of magnitude higher. Table I gives the eigenvalues of the normalized \( X^T X \).

**TABLE I**

Eigenvalues for \( X^T X \) matrix, one year of data

<table>
<thead>
<tr>
<th>( \lambda )</th>
<th>( \lambda_1 )</th>
<th>( \lambda_2 )</th>
<th>( \lambda_3 )</th>
<th>( \lambda_4 )</th>
<th>( \lambda_5 )</th>
<th>( \lambda_6 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \lambda_1 )</td>
<td>3.912</td>
<td>1.309</td>
<td>0.976</td>
<td>0.816</td>
<td>0.480</td>
<td>0.308</td>
</tr>
<tr>
<td>( \lambda_2 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \lambda_3 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The fit achieved with OLS is fairly good, with \( R^2 = 0.6691 \), adjusted \( R^2 = 0.6626 \), and standard error of estimation 485 (compare it to the standard deviation of the dependent variable which is 835). The \( F \)-statistic is high (120.0) corresponding to a very low tail probability, indicating significant regression. The sequential \( \rho \) and the Durbin—Watson statistics do not indicate correlation between the residuals. The OLS estimates and the corresponding \( t \)-statistics of the parameters are given in Table II. The normalized estimates refer to the parameters of the normalized equation. Since the normalized parameters correspond to variates transformed to the same order of magnitude, their comparison indicates the relative contribution of each term, in a deterministic sense, to the dependent variable.

**TABLE II**

**OLS** estimates for one year of data

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimates</th>
<th>Normalized estimates</th>
<th>( t )-statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.35</td>
<td>23,782</td>
<td>11.37</td>
</tr>
<tr>
<td>2</td>
<td>-0.30</td>
<td>-5,830</td>
<td>-3.72</td>
</tr>
<tr>
<td>3</td>
<td>15.18</td>
<td>1,225</td>
<td>2.22</td>
</tr>
<tr>
<td>4</td>
<td>40.27</td>
<td>3,250</td>
<td>4.53</td>
</tr>
<tr>
<td>5</td>
<td>-20.48</td>
<td>-1,672</td>
<td>-2.43</td>
</tr>
<tr>
<td>6</td>
<td>-0.0001</td>
<td>-6,717</td>
<td>-5.45</td>
</tr>
<tr>
<td>7</td>
<td>-0.0065</td>
<td>-916</td>
<td>-0.99</td>
</tr>
<tr>
<td>8</td>
<td>-0.0025</td>
<td>-345</td>
<td>-0.43</td>
</tr>
</tbody>
</table>
Clearly, all parameter estimates are significantly different than zero at the 5% level, except for \( b_7 \) and \( b_8 \). Some of the parameter estimates are highly correlated between themselves, \( b_1 \) and \( b_2 \) have a correlation coefficient of \(-0.877\), \( b_1 \) and \( b_3 \) have a correlation coefficient of \(-0.773\), etc. At this point, we would like to have more information about the stability of the parameter estimates. However, before plotting the ridge trace, we performed minimization of Mallows' \( C_R \) criterion.

The value of the ridge regression coefficient which minimizes \( C_R \) is small:

\[
k_0 = 6.4 \cdot 10^{-4}
\]

For this value of \( k \), the parameter estimates were reduced in absolute value by approximately 1% compared to the OLS solution. The condition number was reduced also by 1% while the residual sum of squares remained practically the same. The square of the Euclidean length of the parameters was reduced by 3%. Most other statistics remained practically unchanged.

Next, the ridge trace was plotted for values of \( k \) between 0 and 0.20 (see Fig. 1). The results are very informative about the stability of parameter estimates. Parameter estimates for \( \beta_1, \beta_2 \) and \( \beta_6 \) are the most unstable of all,

![Normalized Parameters](image)

**Fig. 1.** Ridge trace for one year of data.

dropping in absolute value significantly between \( k = 0 \) and \( k = 0.12 \). The estimate for parameter \( \beta_2 \) is probably the most unstable changing sign at \( k \approx 0.03 \) and then stabilizing at \( b_2 = 2,620 \) for \( k \geq 0.10 \), even though the OLS estimate is \( b_2 = -5,330 \). The estimate for parameter \( \beta_6 \) is more resistant, changing sign at \( k \approx 0.09 \). The parameter estimates for \( \beta_7 \) and \( \beta_8 \) have a behavior which would not be expected by just looking at the OLS results.
These two estimates change sign at \( k = 0.005 \) and stabilize very fast around \( k \approx 0.04 \). As we increase \( k \), they seem to keep their predictive capability. The estimate of parameter \( \beta_5 \) behaves in exactly the opposite way to that of \( \beta_7 \) and \( \beta_8 \). Although it seems significantly different than zero in OLS, it decreases very fast and practically disappears for \( k \geq 0.06 \). Finally parameter estimate \( b_4 \) is very stable throughout the examined range.

We also plotted (Fig. 2) the tradeoff curve of residual sum of squares vs. the square of the Euclidean length of parameter estimates vector. Although the plot does not have any corner that would make our choice easier, it is obvious that we can reduce \( (b^T b)_k \) without getting a significantly worse fit.

![Fig. 2. Tradeoff curve of residual sum of squares vs. square of Euclidean length of normalized parameters, for one year of data.](image)

The ridge trace indicates that all parameter estimates, except probably for \( b_6 \), seem to stabilize around \( k = 0.10 \). This value gives a \( \text{RSS}_k \) which is approximately 12% larger than the fit we obtained in OLS, although \( (b^T b)_k \) was reduced by 80%. The S.E. of estimation increased by approximately 5%. The estimates for \( k = 0.10 \) are given in Table III.

**TABLE III**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Normalized estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.54</td>
<td>9,576</td>
</tr>
<tr>
<td>2</td>
<td>0.14</td>
<td>2,620</td>
</tr>
<tr>
<td>3</td>
<td>20.22</td>
<td>1,632</td>
</tr>
<tr>
<td>4</td>
<td>35.18</td>
<td>2,840</td>
</tr>
<tr>
<td>5</td>
<td>-0.82</td>
<td>-66</td>
</tr>
<tr>
<td>6</td>
<td>0.01</td>
<td>54</td>
</tr>
<tr>
<td>7</td>
<td>0.01</td>
<td>1,401</td>
</tr>
<tr>
<td>8</td>
<td>-0.0078</td>
<td>-1,091</td>
</tr>
</tbody>
</table>
Let us now fit the same model to all available 2828 data points. It turns out that the means, variances and correlations of the variables differ noticeably (and it remains to be seen if significantly) from the previously examined problem. The fit we obtained with the same model is better than in the previous case ($R^2 = 0.7930$, adjusted $R^2 = 0.7925$, standard error of estimation 252). The ordinary least-squares estimates are given in Table IV.

Note that the normalized parameter estimates for each example correspond to different normalizations of $X^TX$.

**TABLE IV**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimates</th>
<th>Normalized estimates</th>
<th>$t$-statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.22</td>
<td>43,719</td>
<td>44.07</td>
</tr>
<tr>
<td>2</td>
<td>-0.183</td>
<td>-6,567</td>
<td>-8.54</td>
</tr>
<tr>
<td>3</td>
<td>5.95</td>
<td>1,649</td>
<td>6.12</td>
</tr>
<tr>
<td>4</td>
<td>13.02</td>
<td>3,611</td>
<td>10.97</td>
</tr>
<tr>
<td>5</td>
<td>-0.37</td>
<td>-2,600</td>
<td>-7.66</td>
</tr>
<tr>
<td>6</td>
<td>-0.0001</td>
<td>-8,326</td>
<td>-18.22</td>
</tr>
<tr>
<td>7</td>
<td>0.0101</td>
<td>2,327</td>
<td>5.92</td>
</tr>
<tr>
<td>8</td>
<td>-0.0087</td>
<td>-2,002</td>
<td>-5.55</td>
</tr>
</tbody>
</table>

It is obvious that using a lot of data (and probably 2828 days of measurements are more data than are available in most commonly encountered cases) does not eliminate the problem of collinearity. From Table IV, it is clear that the OLS estimates of the coefficients, especially $b_1$ and $b_2$, are overinflated, although to a lesser extent than for the case of only 365 days.

Minimizing $C_k$ gave again a very small value of $k$: $k_0 = 6 \cdot 10^{-5}$.

This value of $k$ does not change noticeably the OLS results.

The ridge trace (Fig. 3) is quite similar to the ridge trace for the previously examined case. Estimates $b_1$, $b_2$ and $b_8$ are the least stable ones. This is especially true for $b_2$ which changes dramatically from $b_2 = -6,567$ at $k = 0$, to $b_2 = 0$ at $k \approx 0.018$ and then to $b_2 = 7,083$ at $k = 0.08$. Estimates $b_5$ and $b_6$ tend to go to zero, the former around $k = 0.04$, the latter at $k = 0.10$. Estimates $b_3$ and $b_4$ are stable throughout the range $k = 0$ to $k = 0.20$. Probably one of the interesting results is that parameter estimates $b_7$ and $b_8$, which have the lowest $t$-statistics in OLS, increase in absolute value before they stabilize around $k = 0.04$.

All the parameter estimates are fairly stable around $k = 0.06$. The parameter estimates corresponding to this value of $k$ are given in Table V. For this value of $k$, the RSS is approximately 12% larger than for OLS, while the square of the length of the estimated parameters vector has been reduced by
Fig. 3. Ridge trace for 2828 days of data.

66%. Fig. 4 gives the tradeoff curve of residual sum of square vs. the square of the Euclidian length of the parameter estimates vector.

Comparing the OLS results for the two cases (Tables II and IV), we see considerable differences. For example, the estimate of $\beta_7$ has different signs in each of the cases, while the estimates of $\beta_3$ and $\beta_4$ in one of the cases is approximately three times higher than in the other. Comparing the "stabilized" ridge regression results (Tables III and V), we see that $b_7$ has the same sign

**TABLE V**

Ridge regression estimates for $k = 0.06$, 2828 days of data

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Normalized estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.692</td>
<td>24,765</td>
</tr>
<tr>
<td>2</td>
<td>0.165</td>
<td>5,909</td>
</tr>
<tr>
<td>3</td>
<td>7.408</td>
<td>2,053</td>
</tr>
<tr>
<td>4</td>
<td>12.37</td>
<td>3,430</td>
</tr>
<tr>
<td>5</td>
<td>-0.573</td>
<td>-159</td>
</tr>
<tr>
<td>6</td>
<td>$-0.2538 \times 10^{-4}$</td>
<td>-2,105</td>
</tr>
<tr>
<td>7</td>
<td>0.0193</td>
<td>4,465</td>
</tr>
<tr>
<td>8</td>
<td>-0.0124</td>
<td>-2,862</td>
</tr>
</tbody>
</table>
and approximately the same magnitude in both cases. The estimates of $\beta_3$ and $\beta_4$ still differ significantly, and we must notice that these two parameters are probably suffering the least from collinearity. Consequently, one may suspect that the problem arises from the presence of some nonstationarity in the data.

The ridge trace analysis results for the two cases coincide almost completely with respect to the stability of the estimates: $b_1$, $b_2$ and $b_6$ are unstable, $b_5$ is driven quickly to zero, $b_3$ and $b_4$ are practically unaffected and $b_7$ and $b_8$ tend to strengthen before they stabilize.

STUDYING PARAMETER STABILITY WITH THE KALMAN FILTER

In applying a regression type of approach to the prediction of mean daily river discharges, it is important to know how many years of data are required before the parameter estimates finally stabilize close to what should be their final values. Since the parameter estimate provides the best least-squares fit for the data that have been used, then the stability of the parameter estimates, as more and more data are added, is indicative of the predictive capabilities of the model. The previously discussed ridge regression results implied that this stability may be a far fetched goal.

In order to further test the stability of the parameter estimates, the measurement of mean areal precipitation and river discharge were assumed to be given in a sequential manner. As a new set of measurements is accounted for, the parameters are readjusted. For this on-line estimation, the Kalman filter provides a convenient framework. As dynamic equation, we can consider the constancy of the parameters:

$$\beta(t + 1) = \beta(t)$$ (38)

The measurement equation is:

$$y(t + 1) = \dot X^T(t + 1)\beta(t + 1) + \epsilon(t + 1)$$ (39)
where $\epsilon(t + 1) \sim N(0, \sigma^2)$, while $E\{\epsilon(t_i) \cdot \epsilon(t_j)\} = 0$, $t_i \neq t_j$.

In the Kalman filter, which is the minimum posterior variance estimator, we usually assume that we know the measurement variance $\sigma^2$. Although there are ways to get estimates of this quantity, in our case, we may use an arbitrary estimate of $\sigma^2$, since our purpose is to estimate the parameters and not the error covariance matrix of the estimates. It is convenient to divide the measured quantities $X^T(t)$ and $y(t)$ by $\delta$, so that we obtain a measurement error with variance 1. In this case, Kalman filter provides the best estimate, after $t + 1$ measurements:

$$b(t + 1|t + 1) = b(t|t) + K(t + 1)\{y(t + 1) - X^T(t + 1) b(t|t)\}$$ (40)

where

$$K(t + 1) = \Sigma(t + 1|t) X(t + 1)[1 + X^T(t + 1) \Sigma (t + 1|t) X(t + 1)]^{-1}$$ (41)

For this analysis, the following simple model was used:

$$Q(t) = \beta_1 Q(t - 1) + \beta_2 Q(t - 2) + \beta_3 i(t) + \beta_4 i(t - 1) + \beta_5 i(t - 2)$$ (42)

where $b(t) = \text{best estimate of } \beta \text{ at time } t$; $Q(\cdot) = \text{mean daily flow (ft.}^3\text{ s}^{-1})$; and $i(\cdot) = \text{mean areal precipitation (mm)}$.

As initial conditions, $b(0|0) = 0$ and $\Sigma(0|0) = 10^{20}$ were assumed. This is essentially diffuse prior information about the parameters. The results are given in Table VI.

### TABLE VI

<table>
<thead>
<tr>
<th>Number of observations</th>
<th>$b_1$</th>
<th>$b_2$</th>
<th>$b_3$</th>
<th>$b_4$</th>
<th>$b_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>0.7502</td>
<td>-0.0168</td>
<td>26.24</td>
<td>56.21</td>
<td>-20.24</td>
</tr>
<tr>
<td>400</td>
<td>0.7727</td>
<td>-0.0232</td>
<td>19.78</td>
<td>36.69</td>
<td>-12.16</td>
</tr>
<tr>
<td>600</td>
<td>0.7976</td>
<td>-0.0256</td>
<td>17.87</td>
<td>29.45</td>
<td>-7.35</td>
</tr>
<tr>
<td>800</td>
<td>0.8320</td>
<td>-0.0411</td>
<td>15.26</td>
<td>24.97</td>
<td>-7.20</td>
</tr>
<tr>
<td>1,000</td>
<td>0.8483</td>
<td>-0.0327</td>
<td>10.84</td>
<td>23.15</td>
<td>-8.08</td>
</tr>
<tr>
<td>1,200</td>
<td>0.8701</td>
<td>-0.0411</td>
<td>9.86</td>
<td>23.74</td>
<td>-9.75</td>
</tr>
<tr>
<td>1,400</td>
<td>0.8621</td>
<td>-0.0295</td>
<td>9.34</td>
<td>23.50</td>
<td>-10.85</td>
</tr>
<tr>
<td>1,600</td>
<td>0.8719</td>
<td>-0.0296</td>
<td>10.47</td>
<td>23.10</td>
<td>-11.11</td>
</tr>
<tr>
<td>1,800</td>
<td>0.8730</td>
<td>-0.0319</td>
<td>9.93</td>
<td>22.51</td>
<td>-10.45</td>
</tr>
<tr>
<td>2,000</td>
<td>0.8794</td>
<td>-0.0370</td>
<td>9.40</td>
<td>23.58</td>
<td>-10.36</td>
</tr>
<tr>
<td>2,200</td>
<td>0.8824</td>
<td>-0.0395</td>
<td>8.66</td>
<td>22.89</td>
<td>-9.95</td>
</tr>
<tr>
<td>2,400</td>
<td>0.9004</td>
<td>-0.0460</td>
<td>8.69</td>
<td>22.46</td>
<td>-10.52</td>
</tr>
<tr>
<td>2,600</td>
<td>0.9078</td>
<td>-0.0517</td>
<td>8.41</td>
<td>20.37</td>
<td>-8.88</td>
</tr>
<tr>
<td>2,800</td>
<td>0.9314</td>
<td>-0.0672</td>
<td>8.21</td>
<td>19.60</td>
<td>-9.32</td>
</tr>
</tbody>
</table>
It is noted that even though we fit only 5 parameters, it takes almost 3 years of data until \( b_3, b_4 \) and \( b_5 \) are stabilized near their final values. \( b_1 \) seems to continuously increase even for more than 6 years of data.

Another interesting problem is the numerical accuracy of the Kalman filter. Because of the fact that the Kalman filtering is susceptible to accuracy degradation due to sensitivity to computer roundoff errors, a high-accuracy and stability version of the Kalman filter, known as the Potter algorithm, was used (Potter and Stern, 1963). Additionally, all measurements \( y(t) \) and the elements of \( X(t) \) were normalized so that all numbers were approximately of the same order of magnitude. The final result was checked through a high-accuracy off-line OLS estimation using singular value decomposition and double precision arithmetic. The relative difference was found to be less than \( 10^{-3} \).

However, without conditioning of the measurements to the same order of magnitude, the algorithm failed to converge to the right solution. For 2800 observations the estimates of the unconditioned algorithm were:

\[
\begin{align*}
  b_1 &= 0.7005; &
  b_2 &= 0.0959; &
  b_3 &= 9.15; &
  b_4 &= 21.65 &
  \text{and} \\
  b_5 &= -5.42
\end{align*}
\]

This is indicative of the numerical problems that can appear in Kalman filter estimation and the need to take special measures to improve the numerical accuracy of estimation algorithms.

CONCLUSIONS

It is not easy to decide whether the OLS or the ridge regression parameter estimates are more reasonable, because the parameters do not correspond to a physically based conceptual model. Our model is empirical and is, more or less, evaluated on the basis of the fit it provides to existing data. However, it is known that, in the presence of collinearity, the OLS will go a long way in order to give the best fit. Thus, it gives, for example, \( b_1 \) larger than unity and \( b_2 \) negative. However, the correlation coefficient between the corresponding carrier \( Q(t-1) \) and \( Q(t-2) \) is 0.851, so that as noted by Hoerl and Kennard (1970b), they are almost the same variables with a different name. It would be rather surprising if the effect of \( Q(t-1) \) on \( Q(t) \) is positive, while the effect of \( Q(t-2) \) on \( Q(t) \) is negative.

Nevertheless ridge regression seems to provide parameter estimates that seem more intuitive when seen in light of the above argument. It handles collinearity problems in a consistent objective manner and without endangering regression accuracy. The OLS and ridge regression parameter estimates should be tested in the predictive mode.

The plotting of the ridge trace and the RSS vs. length tradeoff curves provide valuable information about parameter stability in the presence of collinearity. Ridge regression is a valuable sensitivity analysis tool.
Both the ridge regression and Kalman filter studies indicated the need of large amounts of data to obtain reliable stable parameters. Such characteristics are essential in the prediction framework.

Finally the numerical problems related to common Kalman filter algorithms were illustrated. The user of such techniques must be aware of the sensitivity of his result to numerical errors.

ACKNOWLEDGMENTS

This work was sponsored by the Department of Commerce, National Weather Service, Hydrologic Research Laboratory. It is part of a larger study titled On-line river discharge forecasting using filtering and estimation theory, Contract No. 7-35112. Work was performed at the Ralph M. Parsons Laboratory for Water Resources and Hydrodynamics, Department of Civil Engineering, M.I.T., Cambridge, Mass.

This paper was presented at the American Geophysical Union Chapman Conference on Kalman Filter Applications of Kalman Filter to Hydrology, Hydraulics and Water Resources, Pittsburgh, Pa., May 1978.

REFERENCES