J. R. Statist. Soc. B (1996)
58, *No.* 4, *pp.* 703–710

Continuum Regression is not Always Continuous

By ANDERS BJÖRKSTRÖM and ROLF SUNDBERG†

Stockholm University, Sweden

[Received July 1995. Revised November 1995]

SUMMARY

Continuum regression (CR) is regression of a response y on that linear combination t_γ of explanatory variables which maximizes $r^2(y, t)\,\mathrm{var}(t)^\gamma$, where γ is chosen from a continuum of candidates. CR includes several common methods as special cases: ordinary least squares, principal component regression, partial least squares and a modified ridge regression. We demonstrate that CR, despite its name, can yield a predictor that is a discontinuous function of γ and of the calibration data. We illustrate this with a set of real data. The relationship between CR and ridge regression is the key to understanding this phenomenon.

Keywords: CROSS-VALIDATION; REGRESSION METHODS; RIDGE REGRESSION

1. INTRODUCTION

In regression problems where the explanatory variables are nearly collinear, the ordinary least squares (OLS) estimator has an unpleasantly large variance. Several so-called regularized (or shrinkage) methods, trading variance for bias, have been suggested as alternatives to OLS. Well-known examples are principal component regression (PCR), partial least squares (PLS) regression and ridge regression (RR); see Brown (1993) or Frank and Friedman (1993) for reviews. Stone and Brooks (1990) introduced the method of 'continuum regression' (CR), in which they considered a spectrum of possible regressors, each associated with a value of a parameter γ. An appealing feature of this technique is that it incorporates OLS, PLS and PCR as special cases, corresponding to particular values of γ, and also a modified RR method.

Following the customary notation, we denote by X an $n \times p$ design matrix and by y the vector of n responses, such that the ith row of X gives the values of the p explanatory variables for the ith observation, and y_i the corresponding value of the response variable. The vector y and all columns of X are assumed centred. The problem, generally speaking, is to find a p-vector b such that $b^{\mathrm{T}}x$ makes a good predictor for known x. In OLS as well as in first-factor PLS or PCR the vector b is determined by simple linear regression of y on a one-dimensional regressor Xc, where the coefficient vector c is chosen by the following criteria for OLS, PLS and PCR respectively:

(a) the squared correlation $r^2(y, Xc)$ between y and Xc is maximized;
(b) the covariance between y and Xc is maximized under the constraint $|c| = 1$;
(c) the sample variance of Xc (or squared length $|Xc|^2$) is maximized, also under $|c| = 1$.

†*Address for correspondence*: Institute of Actuarial Mathematics and Mathematical Statistics, Stockholm University, S-10691, Stockholm, Sweden.
E-mail: rolfs@matematik.su.se

 0035–9246/96/58703

This analogy between OLS, PLS and PCR was pointed out by Stone and Brooks (1990), who went on to formulate their principle of CR, where a vector c with $|c| = 1$ is chosen for each $\gamma \geqslant 0$ by maximizing

$$T(\gamma, c) = (y^{\mathrm{T}} Xc)^2 \, |Xc|^{2(\gamma-1)} \propto r^2(y, Xc) \, |Xc|^{2\gamma}.$$

In this representation, $\gamma = 0$ corresponds to OLS, $\gamma = 1$ to PLS and $\gamma \to \infty$ yields PCR.

Having found this c, $c = c_\gamma$ say, we construct a predictor $b_\gamma^{\mathrm{T}} x$ by simple linear regression of y on Xc_γ, i.e. we multiply $c_\gamma^{\mathrm{T}} x$ by the corresponding regression coefficient. Stone and Brooks (1990) advocated that the choice of γ be made by cross-validated optimization over the predictors $b_\gamma^{\mathrm{T}} x$. The procedure can be repeated using the residuals as input, yielding an adjusted estimator, but we shall not consider more than the first step in the sequence, i.e. first-factor CR.

RR is equivalent to selecting as b the vector that minimizes $|Xb - y|^2 + \delta |b|^2$ where $\delta \geqslant 0$ is an adjustable parameter. The relationship between RR and CR was demonstrated by Sundberg (1993): CR differs from RR only by a scalar factor, that should be chosen to minimize the residual sum of squares, and any CR regressor that corresponds to a value of γ between 0 and 1 is in fact a ridge regressor thus modified. If we allow $\delta < 0$ this correspondence can be extended to CR for $\gamma > 1$, as remarked by de Jong and Farebrother (1994).

Finally, it should be mentioned that Brooks and Stone (1994) have also developed a multivariate version of CR.

2. DISCONTINUITIES IN CONTINUUM REGRESSION

In our studies of CR, we have sometimes observed the predictor obtained by CR to change discontinuously as the CR parameter γ is varied (or, when data vary). In retrospect, it is not difficult to see the cause of this. First-factor CR involves the maximization of a function:

$$c_\gamma \equiv \arg \max_{|c|=1} \{T(\gamma, c)\}$$

and, even though T is continuous in both c and γ, the argument c_γ thus defined need not be continuous in γ. Fig. 1 shows in principle what happens, by an example with $p = 2$. Fig. 1 shows $T(\gamma, c)$ as a function of c for $\gamma = 0.6$ and $\gamma = 0.7$. The point to note is that, as γ is changed, one of the two local maxima is reduced and the other is increased, taking over as global maximum. Consequently, there is a jump for some value of γ between 0.6 and 0.7. Similarly, γ could have been held constant at 0.7 while the curves for $(X^{\mathrm{T}} y) = (3, 1)$ and $(X^{\mathrm{T}} y) = (2.5, 1)$ were compared (simulating a change in the response data y). This pair of curves would also have demonstrated the existence of a jump, if plotted.

It is uncomfortable only to know that a predictor might make a sudden jump in response to a slight change in γ. Therefore, we now investigate under what conditions a data set is capable of producing this effect.

As shown in Sundberg (1993), first-factor CR is related to RR in the following way. The CR predictor is proportional to an RR predictor

$$c_\gamma \propto b^{\mathrm{RR}}(\delta),$$

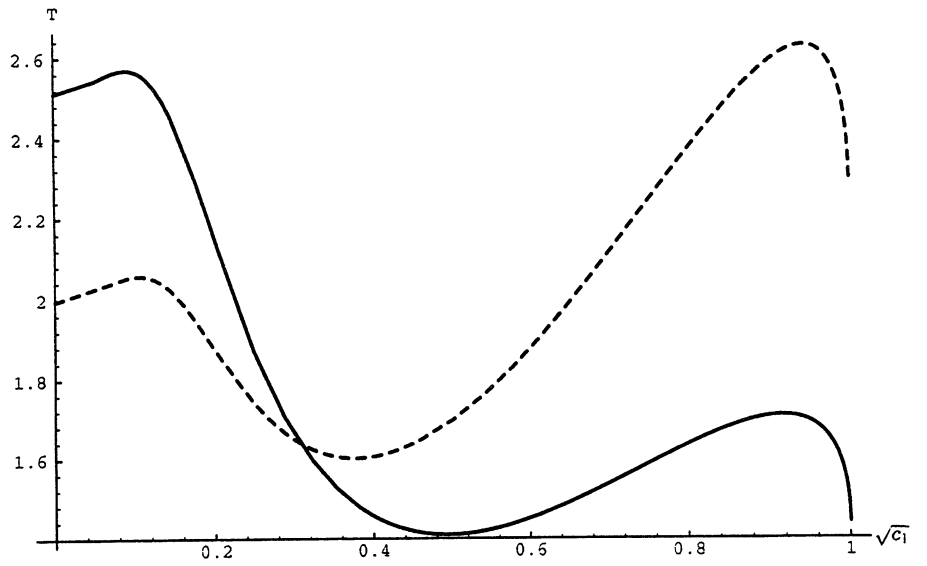

Fig. 1. Illustration of a function $T(\gamma, c)$ to be maximized, as a function of c, for $\gamma = 0.6$ (———) and $\gamma = 0.7$ (- - - -): the data have $p = 2$, with $X^T y = (3, 1)^T$ and $X^T X = \text{diag}(100, 0.01)$ (on the abscissa, c is represented by the square root of its first co-ordinate c_1)

where $\delta = \delta_{CR}(\gamma)$ is a monotone function of γ. In particular, $\delta_{CR}(0) = 0$ (OLS), $\delta_{CR}(1) = \pm\infty$ (PLS) and $\delta_{CR}(\infty) = -\lambda_1$ (PCR: λ_1 is the largest eigenvalue of $X^T X$).

By definition, $b^{RR}(\delta) = (X^T X + \delta I)^{-1} X^T y$. Thus, $b^{RR}(\delta)$ is a continuous function of δ, and so is its projection on the unit sphere. Discontinuities in CR can occur because the optimal value of δ may change discontinuously with γ, i.e. $\delta_{CR}(\gamma)$ may be a discontinuous function.

Equation (2.4) in Sundberg (1993) states, in our notation, that

$$\delta = \frac{\gamma}{1-\gamma} \frac{|Xc_\gamma|^2}{|c_\gamma|^2}. \tag{1}$$

Since this is a scale invariant formulation, we can replace c_γ with $b^{RR}(\delta)$. It is convenient to express vectors and matrices in the canonical basis of right singular vectors of X. We perform a singular value decomposition $X = USV^T$ and write $X^T y = Vg$, $c = Vz$ and $S^T S = \Lambda$. Then equation (1) reads

$$\delta = \frac{\gamma}{1-\gamma} \frac{|\Lambda^{1/2} (\Lambda + \delta I)^{-1} g|^2}{|(\Lambda + \delta I)^{-1} g|^2}.$$

This can be written

$$\delta = \frac{\gamma}{1-\gamma} \frac{P(\delta)}{Q(\delta)} \tag{2}$$

where $P(\delta)$ and $Q(\delta)$ are two positive definite polynomials of degree $2(p-1)$:

$$P(\delta) = \sum_{j=1}^{p} g_j^2 \lambda_j \prod_{k \neq j} (\lambda_k + \delta)^2,$$

$$Q(\delta) = \sum_{j=1}^{p} g_j^2 \prod_{k \neq j} (\lambda_k + \delta)^2.$$

When equation (2) is solved for γ, we obtain

$$\gamma = \gamma(\delta) = \frac{\delta Q(\delta)}{P(\delta) + \delta Q(\delta)}. \tag{3}$$

For given γ, equation (2) holds for $\delta = \delta_{CR}(\gamma)$, but it may hold for other δ as well. The function $\delta_{CR}(\gamma)$ is uniquely obtained from equation (3) by inversion precisely when $\gamma(\delta)$ in equation (3) is a monotone function of δ. Generally, for a given γ, several δ may solve equation (3). Fig. 2 illustrates this. (See Section 3 for more details on this data set.) For $0.580 \leqslant \gamma_0 \leqslant 0.593$, the equation $\gamma(\delta) = \gamma_0$ will have three solutions. The existence of more than one solution does not make the CR regressor ambiguous but only means that the global maximum of $T(\gamma, c)$ on the sphere $|c| = 1$ is to be found among a finite number of candidates. However, it is easily realized from Fig. 2 that $\delta_{CR}(\gamma)$ *must* have discontinuities in such cases. Conversely, it is evident that $\delta_{CR}(\gamma)$ lacks discontinuities when equation (3) is everywhere increasing.

In Fig. 2 we are between OLS ($\gamma = 0$) and PLS ($\gamma = 1$), but the above argument is also valid for $\gamma > 1$. When $\gamma > 1$, the condition for no jumps is that equation (3) must map the interval $\delta \in (-\infty, -\lambda_1)$ on $\gamma \in (1, \infty)$ in a monotone way.

In summary, we have seen that a necessary and sufficient condition for a data set to be free from jumps in the CR predictors is that the eigenvalues $\lambda_1, \ldots, \lambda_p$ and the coefficients $g_1, \ldots, g_p$ for $X^T y$ in the basis of eigenvectors are such that the function

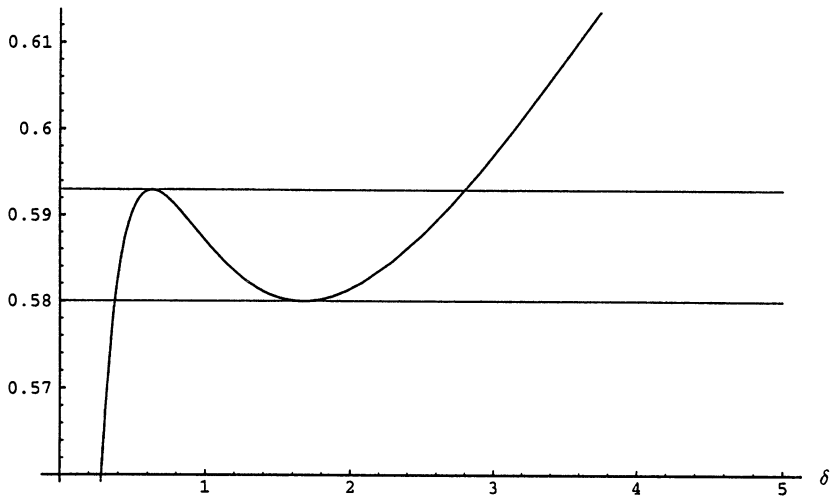

Fig. 2. Illustration of formula (3), i.e. of the CR parameter γ regarded as the function $\gamma(\delta)$ of the RR parameter δ (the data are the first nine observations of Fearn (1983))

$\delta Q(\delta)/\{P(\delta) + \delta Q(\delta)\}$ is monotone or, equivalently, that $\delta Q(\delta)/P(\delta)$ is an increasing function for all δ. The derivative of $\delta Q(\delta)/P(\delta)$ is $(PQ + \delta PQ' - \delta P'Q)/P^2$. The function $\gamma(\delta)$ is monotone if and only if the equation

$$PQ + \delta PQ' - \delta P'Q = 0 \tag{4}$$

has no solutions in the region $\{-\infty, -\lambda_1\} \cup \{0, \infty\}$. This offers a way to explore whether a given data set will yield discontinuities.

There is not likely to be a simple answer to the question which combinations of $\lambda_1, g_1, \ldots, \lambda_p, g_p$ yield or do not yield discontinuities. By experience, when $p = 2$ we can say qualitatively that for a fixed ratio g_1/g_2 a discontinuity will appear if the ratio λ_1/λ_2 is sufficiently large, whereas for a fixed ratio λ_1/λ_2 the discontinuity disappears if the ratio g_1/g_2 is sufficiently extreme (large or small).

2.1. *Remark on Cross-validation*

Stone and Brooks (1990) used a 'leave-one-out' cross-validatory index I_γ for the choice of γ. In essence, I_γ is a linear function of $\Sigma (\hat{y}_{/i} - y_i)^2$, where $\hat{y}_{/i}$ is the predictor of y_i based on the other $n - 1$ observations. If a discontinuity is present in any of the n reduced data sets, it will cause a discontinuity in the corresponding term $(\hat{y}_{/i} - y_i)^2$ as a function of γ, and as a consequence also in the function I_γ.

3. EXAMPLE: FEARN'S NEAR INFRA-RED DATA FOR CALIBRATION OF PROTEIN

Here we refer to Fearn's (1983) Table 1, now classical protein data, to demonstrate that the occurrence of jumps is not an unusual fabricated phenomenon. Among many others, Stone and Brooks (1990), example 3, have used these data, when they illustrated their CR methodology on the first 12 items of Fearn's Table 1. Here we reduce the data set slightly more, to consist of the first nine or 10 items only. There are $p = 6$ explanatory variables, representing the log(1/reflectance) values at six different wavelengths and used to predict the percentage y of protein in wheat samples. Like in Fearn (1983), we standardized the explanatory variables but attempted no other transformation. Solving equation (4) for the first $n = 9$ data items, we found that the function $\gamma(\delta)$ had a local maximum and a local minimum on the positive real axis. Fig. 2 shows the graph of $\gamma(\delta)$ in the region around $\gamma = 0.6$. For γ between 0.580 and 0.593 equation (3) has three solutions in δ. In Fig. 3 we illustrate the maximum $T\{\gamma(\delta), c_{\gamma(\delta)}\}$ for $0.3 \leqslant \delta \leqslant 3.3$.

To illustrate what implications the discontinuity has for cross-validation, we augment the data set with Fearn's 10th observation and show the 'leave-one-out' index I_γ in Fig. 4. We note a downward jump at $\gamma \approx 0.588$, when the continuum regressor turns over from the descending to the ascending segment in Fig. 3. The cross-validation index drops instantly from 0.515 to 0.353. The prediction $\hat{y}_{/10}$ for the left-out observation drops from 9.89 to 8.93. (The value actually observed for y_{10} was 11.39, so both predictions are too low, and it is clear from Fig. 4 that a CR parameter of $\gamma = 0.588$ is not near the range of values that one would use in practice for these data.)

What happens with the CR predictor as γ is varied? As γ approaches 0.588 the CR predictor successively distributes more weight to the first principal component,

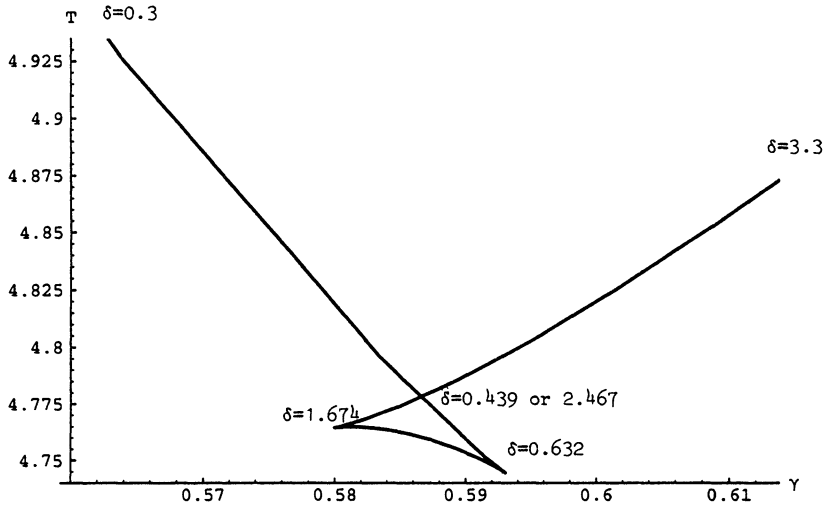

Fig. 3. Illustration of the maximum $T\{\gamma(\delta), c_{\gamma(\delta)}\}$ as a function of $\gamma(\delta)$, when δ runs from 0.3 to 3.3 (the data are the first nine observations of Fearn (1983))

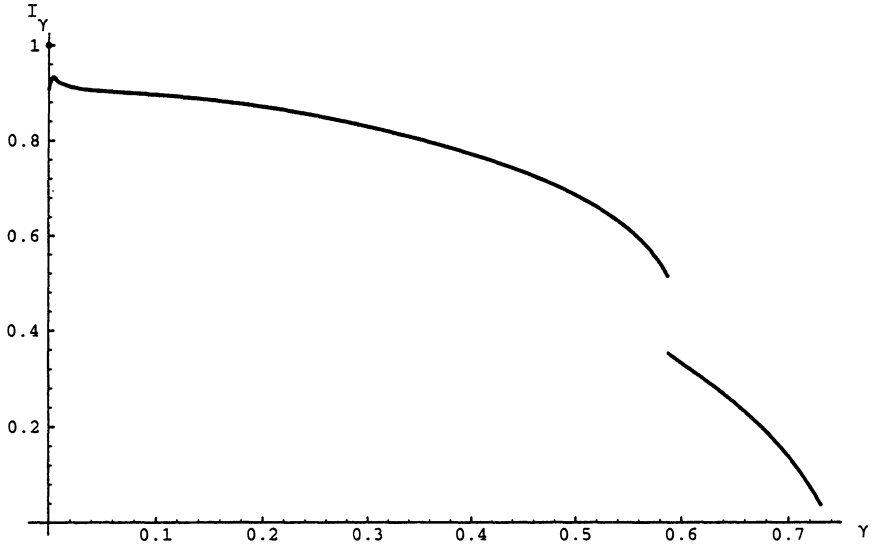

Fig. 4. Cross-validatory index I_γ as a function of γ for first-factor CR, using as data the first 10 observations of Fearn (1983)

whose regression coefficient is increased from 0.46 (at OLS) to 1.02, whereas the others are more or less shrunk. At the jump point even the first coefficient drops, to 0.77. From there it continues to decrease with increasing γ, like the others (the ultimate value for PCR being 0.46 again, of course). This drop of all coefficients at the discontinuity could be expected as a consequence of the proposition in Sundberg (1993).

4. CONSEQUENCES FOR CONTINUUM REGRESSION IN PRACTICE

The two curves in Fig. 1 were drawn for the same data set X, y and different values of γ. It is easily realized that the same effect (one local maximum growing larger than another) can also arise if γ is held constant and some items of the data set are varied instead. Thus, for fixed γ, two data sets which differ only by some insignificant amount in some variable in one observation may produce radically different CR predictions. However, we see no reason why this should be a serious shortcoming of CR in practice. In fact, several other model selection techniques share this property. Our main worry, instead, concerns the possibility of devising efficient algorithms for CR. Our present method to find a continuum regressor is to compute all real solutions to equation (3) (using subroutine 'NSolve' in MATHEMATICA, version 2.2.2 (Wolfram Research, 1992)). If there is more than one solution, we find the global maximum by comparing $T(\gamma, c_\gamma)$ for each of them. This is a slow procedure. The subset of Fearn's (1983) data ($n=9$, $p=6$) required about 150 s to compute one cross-validation index on a Macintosh Power 6100 computer. To apply CR to large data sets, more efficient algorithms will be needed. (This is even more essential when we want to allow more than one factor.) It seemed to us that a promising way would be to use iteration of equation (2):

$$\delta_{n+1} = \frac{\gamma}{1-\gamma}\frac{P(\delta_n)}{Q(\delta_n)}.$$

However, the possibility of discontinuities poses a non-trivial problem: when equation (3) for δ has more than one solution, the iteration just mentioned will have more than one stationary point, i.e. $c_{\gamma(\delta_n)}$ will not converge to the appropriate regressor for all choices of starting point δ_0. This motivates considering δ as a more basic parameter than γ. We address this question in Björkström and Sundberg (1996).

5. CONCLUSIONS

For some data sets, the CR regressor coefficients vector c_γ is a discontinuous function of the CR parameter γ. In such cases, the correspondence between CR regressors and RR predictors (Sundberg, 1993) is not one to one; instead, the CR regressors form a subset of the RR predictors. When the cross-validatory index I_γ is plotted as a function of γ, discontinuities are observed as jumps in the graph.

REFERENCES

Björkström, A. and Sundberg, R. (1996) A generalized view on continuum regression. *Report*. Institute of Actuarial Mathematics and Mathematical Statistics, Stockholm University, Stockholm.

Brooks, R. and Stone, M. (1994) Joint continuum regression for multiple predictands. *J. Am. Statist. Ass.*, **89**, 1374–1377.

Brown, P. J. (1993) *Measurement, Regression, and Calibration*. Oxford: Oxford University Press.

Fearn, T. (1983) A misuse of ridge regression in the calibration of a near infrared reflectance instrument. *Appl. Statist.*, **32**, 73–79.

Frank, I. E. and Friedman, J. H. (1993) A statistical view of some chemometrics regression tools. *Technometrics*, **35**, 109–148.

de Jong, S. and Farebrother, R. W. (1994) Extending the relationship between ridge regression and continuum regression. *Chemometr. Intell. Lab. Syst.*, **25**, 179–181.

Stone, M. and Brooks, R. J. (1990) Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression (with discussion). *J. R. Statist. Soc.* B, **52**, 237–269; correction, **54** (1992), 906–907.

Sundberg, R. (1993) Continuum regression and ridge regression. *J. R. Statist. Soc.* B, **55**, 653–659.

Wolfram Research (1992) *Mathematica — a System for Doing Mathematics by Computer*. Champaigne: Wolfram Research.