ELSEVIER

Chemometrics and Intelligent Laboratory Systems 25 (1994) 179–181

Chemometrics and intelligent laboratory systems

Extending the relationship between ridge regression and continuum regression

Sijmen de Jong [a,*], Richard W. Farebrother [b]

[a] *Unilever Research Laboratorium Vlaardingen, P.O. Box 114, 3130 AC Vlaardingen, The Netherlands*

[b] *Department of Econometrics and Social Statistics, Victoria University of Manchester, Manchester M13 9PL, United Kingdom*

Received 16 March 1994; accepted 11 May 1994

Abstract

Recently, a close relationship has been established between ridge regression (RR) and a special case of continuum regression (CR). Attention was restricted to the usual positive range of values for the ridge parameter. This restriction identifies the trajectory lying between ordinary least squares (OLS) and partial least squares (PLS) regressions, leaving the trajectory between PLS and principal component regression (PCR) untouched. In this note we demonstrate that the relationship between CR and RR can be extended to the full range of methods, OLS ↔ PLS ↔ PCR, identified by the CR technique. For this purpose one has to admit a nonstandard variant of the RR technique in which the ridge parameter becomes negative.

1. Introduction

Ridge regression (RR), partial least squares (PLS), and principal components regression (PCR) are three of the more familiar alternatives to ordinary least squares (OLS) regression when one is concerned with the problem of estimating the slope parameters of the standard linear model in the presence of highly correlated predictor variables [1–3]. Stone and Brooks [4] have introduced a new fitting technique called continuum regression (CR) which contains three of the above four methods as particular cases. In a natural parameterization of this technique, the continuum of the title is represented by a unit interval with a point corresponding to OLS at one end ($\alpha = 0$), PLS in the middle ($\alpha = 0.5$), and PCR at the other end ($\alpha = 1$). This process of unification was continued by Sundberg [5] who showed that there is a close relationship between RR and a special case of CR. More precisely, he established that the first factor continuum regression (FFCR) estimator is a scalar multiple of the standard RR estimator. Sundberg restricted attention to the usual range $\delta \geq 0$ for the ridge parameter δ, which corresponds to the range $0 \leq \alpha < 0.5$ for the natural parameter, or to $0 \leq \gamma < 1$ for the CR tuning parameter $\gamma = \alpha/(1-\alpha)$. In view of the fact that the CR estimator is defined for all nonnegative values of the tuning parameter γ, it seems unnecessary to restrict its range in this way. Indeed we will now establish that this restriction is not imperative if one is willing to

* Corresponding author.

SSDI 0169-7439(94)00044-J

admit a nonstandard variant of the RR technique.

2. Theory

Consider the standard linear model $y = X\beta + \varepsilon$, where y is the $n \times 1$ vector of responses, X the $n \times p$ matrix of regressors, β the $p \times 1$ vector of regression parameters, and ε the $n \times 1$ vector of errors. The FFCR variant of Stone and Brooks's procedure first determines the $p \times 1$ direction vector c of the estimated slope parameters $b = mc$ by minimizing the CR criterion

$$T = \frac{c'ss'c}{c'Sc}\left(\frac{c'Sc}{c'c}\right)^{\gamma} \tag{1}$$

where $S = X'X$, $s = X'y$ and $\gamma \geq 0$. Given a suitably normalized value for c, the scalar parameter m follows from the least squares regression of y on Xc, namely $m = c's/(c'Sc)$.

Sundberg showed that the CR solution corresponds to that of ridge regression:

$$c \propto (S + \delta I)^{-1}s \tag{2}$$

when the ridge parameter δ is chosen as

$$\delta = [\gamma/(1-\gamma)] \cdot (c'Sc)/(c'c) \tag{3}$$

Sundberg only considered the interval $0 \leq \gamma < 1$ which corresponds to the usual range of nonnegative values for the ridge parameter. For $\gamma = 0$ one has $\delta = 0$ and, using (2), $c \propto S^{-1}s$, which is proportional to the OLS solution for b. As $\gamma \uparrow 1$ one finds $\delta \uparrow +\infty$. In this limiting case one may write $c \propto (S + \delta I)^{-1}s$ or $c \propto (S/\delta + I)^{-1}s$. As S/δ approaches the $p \times p$ null matrix for $\delta \uparrow +\infty$ one obtains $c \propto I^{-1}s$ or $c \propto s$ as a limiting solution. Now the unnormalized weight vector w_1 defining the first component in PLS regression is also given by $s = X'y$. Thus the direction vectors in ridge regression for $\delta \uparrow +\infty$, in first-factor continuum regression for $\gamma \uparrow 1$, and in first-factor PLS regression coincide. Of course, the equivalence of PLS regression and CR for $\gamma = 1$ is a necessary consequence from the equivalence of their optimality criteria for $\gamma = 1$ [4]. The interpretation of first-factor PLS as an unshrunken ridge regression solution in the limit as $\delta \uparrow +\infty$ has been given before by De Jong and Kiers [6].

We now extend Sundberg's analysis of the FFCR problem by considering the range $1 \leq \gamma < +\infty$. δ is not defined for the value $\gamma = 1$. However, if γ approaches the value 1 from above, $\gamma \downarrow 1$, then $\delta \downarrow -\infty$ and the solution $c \propto -s$ for $\gamma \downarrow 1$ only differs in sign from the solution $c \propto s$ for $\gamma \uparrow 1$. The scalar m allows for this sign difference and the complete regression vectors, $b = mc$, are equal for the two limiting cases $\gamma \downarrow 1$ and $\gamma \uparrow 1$. Hence there is no discontinuity in the solution vectors at the value $\gamma = 1$ and the name 'continuum regression' is still appropriate.

To explore the region $\gamma > 1$ it is expedient to perform an orthogonal rotation to the columns of X. Let the $p \times p$ positive definite matrix $S = X'X$ have eigenvalue decomposition $S = V\Lambda V'$, where Λ is a $p \times p$ diagonal matrix of eigenvalues arranged in weakly descending order $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ and V is the corresponding $p \times p$ matrix of orthonormal eigenvectors. (Notice that we use the usual descending ordering of eigenvalues here. This is opposite to Sundberg's choice [5].) In this context let $g = V's = V'X'y$ and $f = V'c$. One then finds that:

$$f_i \propto g_i/(\lambda_i + \delta), \quad i = 1,2,\ldots,p \tag{4a}$$

or, equivalently,

$$f_i \propto g_i/(\lambda_i/\delta + 1), \quad i = 1,2,\ldots,p \tag{4b}$$

and

$$\delta = [\gamma/(1-\gamma)]\Sigma f_i^2\lambda_i/\Sigma f_i^2 \tag{5}$$

The term $\Sigma f_i^2\lambda_i/\Sigma f_i^2$ represents a weighted average of the eigenvalues of S, say $\tilde{\lambda}$, hence it can only take finite positive values lying between the minor eigenvalue λ_p and the major eigenvalue λ_1. Eq. (5) reveals that δ increases from 0 to $+\infty$ as γ increases from 0 to 1, as we have already seen. At the same time the elements of f are continuously redistributed from $f \propto \Lambda^{-1}g$ (OLS) to $f \propto g$ (PLS) as δ increases from 0 to $+\infty$ (see Eqs. 4a and 4b, respectively). In a similar way, as γ increases from 1 to $+\infty$, Eq. 5 shows that δ increases from $-\infty$ to $-\tilde{\lambda}$. When $\delta \downarrow -\infty$ the weighting coefficients $1/(\lambda_i/\delta + 1)$ applied to the elements of g are all equal to 1 and $f \propto g$ (see

Eq. 4b). As δ increases more weight is given to the eigenvectors corresponding to the larger eigenvalues since, for $\delta < -\lambda_1$, $i<j \Rightarrow \lambda_i \geq \lambda_j \Rightarrow \lambda_i/\delta \leq \lambda_j/\delta \Rightarrow \lambda_i/\delta + 1 \leq \lambda_j/\delta + 1 \Rightarrow 1/(\lambda_i/\delta + 1) \geq 1/(\lambda_j/\delta + 1)$. In the limiting case $\delta \uparrow -\lambda_1$ we find that $1/(\lambda_1/\delta + 1) \uparrow +\infty$, whereas $1/(\lambda_j/\delta + 1)$ remains finite for $j > 1$ provided $\lambda_1 > \lambda_2$. (If λ_1 is not strictly greater than λ_2 then the situation becomes too complex to be briefly summarized here.) Thus, all the weight is placed on the first element of $\boldsymbol{g}$ giving $\boldsymbol{f} = (1,0,0,\ldots,0)'$. One may readily verify that $\delta = -\lambda_1$ and $\boldsymbol{f} = (1,0,0,\ldots,0)'$ indeed forms a solution of Eqs. (4) and (5) for $\gamma \uparrow +\infty$. Premultiplying $\boldsymbol{f} = (1,0,0,\ldots,0)'$ by the $p \times p$ matrix $\mathbf{V}$, we find that this limiting solution corresponds to $\boldsymbol{c} \propto \boldsymbol{v}_1$, that is, to the eigenvector which defines the dominant principal component, as in first-factor PCR.

In summary our results and those of Sundberg imply a continuous remodelling which passes from $\boldsymbol{c} \propto \mathbf{S}^{-1}\boldsymbol{s}$, the OLS solution for $\gamma = \delta = 0$, via $\boldsymbol{c} \propto \boldsymbol{s}$, the PLS solution for $\gamma = 1$ or $\delta = \pm\infty$, to $\boldsymbol{c} \propto \boldsymbol{v}_1$, the PCR solution for $\gamma = +\infty$ or $\delta = -\lambda_1$ as the natural parameter increases from 0 to 1. As an immediate consequence of these results we find that δ is necessarily nonnegative when $0 \leq \gamma < 1$ so that $\mathbf{S} + \delta\mathbf{I}$ is a positive definite matrix and the FFCR estimator is proportional to the standard RR estimator $(\mathbf{S} + \delta\mathbf{I})^{-1}\boldsymbol{s}$. Similarly, when $\gamma > 1$, δ is necessarily less than $-\lambda_1$ so that $\mathbf{S} + \delta\mathbf{I}$ becomes a negative definite matrix. Division by the negative ridge constant δ gives $\delta^{-1}\mathbf{S} + \mathbf{I}$ as a positive definite matrix, and we may deduce that the FFCR estimator is proportional to the nonstandard RR estimator $(\delta^{-1}\mathbf{S} + \mathbf{I})^{-1}\boldsymbol{s}$ for the range $1 < \gamma < +\infty$.

Table 1
Equivalence of first-factor continuum regression and rescaled ridge regression with other biased regression methods

Method	Criterion	α	γ	δ	$\boldsymbol{c}$
(FF-)OLS	Correlation	0	0	0	$\propto \mathbf{S}^{-1}\boldsymbol{s}$
FF-PLS	Covariance	0.5	1	$\pm\infty$	$\propto \boldsymbol{s}$
FF-PCR	Variance	1	∞	$-\lambda_1$	$\propto \boldsymbol{v}_1$

3. Conclusion

Sundberg's characterization of the FFCR estimator [5] may readily be generalized to all nonnegative values of the tuning parameter γ, thus covering the full range of methods, OLS $\leftrightarrow$ PLS $\leftrightarrow$ PCR, implied by this technique. In principle we may extend the analysis to negative values of γ, but this second generalization is unhelpful as it corresponds to our favouring low-variance directions for the parameter estimates.

Table 1 summarizes the relations between the various parameters and conditions.

References

[1] A.E. Hoerl and R.W. Kennard, Ridge regression: biased estimation for nonorthogonal problems, *Technometrics*, 12 (1970) 55–67.

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

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

[4] M. Stone and R.J. Brooks, Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components segression (with discussion), *Journal of the Royal Statistical Society, Series B*, 52 (1990) 237–269.

[5] R. Sundberg, Continuum regression and ridge regression, *Journal of the Royal Statistical Society, Series B*, 55 (1993) 653–659.

[6] S. de Jong and H.A.L. Kiers, Principal covariates regression. Part I: Theory, *Chemometrics and Intelligent Laboratory Systems*, 14 (1992) 155–164.