© Board of the Foundation of the Scandinavian Journal of Statistics 1999. Published by Blackwell Publishers Ltd, 108 Cowley Road, Oxford OX4 1JF, UK and 350 Main Street, Malden, MA 02148, USA Vol 26: 17–30, 1999

A Generalized View on Continuum Regression

ANDERS BJÖRKSTRÖM and ROLF SUNDBERG

Stockholm University, Sweden

ABSTRACT. We generalize the relationship between continuum regression (Stone & Brooks, 1990) and ridge regression, by showing that any optimization principle will yield a regressor proportional to a ridge regressor, provided only that the principle implies maximizing a function of the regressor's sample correlation coefficient and its sample variance. This relationship shows that continuum regression as defined via ridge regression ("least squares ridge regression") is a more generally valid methodology than previously realized, and also opens up for alternative choices of its second and subsequent factors.

Key words: cross-validation, linear prediction, near collinearity, partial least squares regression (PLS), principal component regression (PCR), ridge regression, shrinkage

1. Introduction

1.1. The problem

In regression problems where the explanatory variables are near collinear, the ordinary least squares (OLS) predictor, albeit unbiased, has large variance. This often leads to large and unrealistic predictor coefficients. Several so-called shrinkage or regularized regression methods, trading bias for variance, have therefore been proposed. Well-known methods are principal component regression (PCR), partial least squares regression (PLS), ridge regression (RR), and continuum regression (CR); see Stone & Brooks (1990), Frank & Friedman (1993), Brown (1993) or Sundberg (1999) for reviews. Most of these methods are applied in steps: the residuals from one step are taken as inputs for the next step, and an additional regressor is constructed. Cross-validation is usually used to choose an optimal number of steps (factors).

Questions about interrelationships between the various regularized methods have both practical importance and theoretical interest, as has the question under what conditions higher factors are worthwhile to compute. Of special interest here is CR, being wide enough to include OLS, PLS and PCR as special cases.

Sundberg (1993) demonstrated that first-factor CR may be interpreted as a modified form of RR, the modification consisting of a scale adjustment of the predictor in order to minimize the residual sum of squares over the training data. In the present paper, we demonstrate that this modified form of RR is a very general method, in the sense that a wide class of criteria in fact yields predictors that are ridge regressors modified for scale. A sufficient condition is that the method be equivalent to maximizing a function of the correlation and variance in data. To give just one example, it is well-known that first-factor PLS regression is equivalent to using the regressor that maximizes sample covariance between the response values and their fitted values. Since covariance is the product of correlation and variance, the first-factor PLS regressor is proportional to a ridge regressor (more precisely, in the limit as the ridge constant tends to infinity).

Based on this general result, we suggest a generalized formulation of continuum regression which is free from a type of discontinuity that may appear with the original CR (Björkström &

Sundberg, 1996). We also demonstrate that for this generalized formulation, the coefficient of determination R^2 and the length of the estimator are decreasing functions of the continuum parameter, for the first factor.

1.2. Notations, terminology and basic assumptions

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, so that the rank of X can be at most $\min(n-1, p)$.

A *linear predictor* is, strictly speaking, an expression $b^{\mathrm{T}}x$ which is used for prediction of y, where (x, y) is an item of data with y not yet observed. However, we will often call the p-vector b itself a "predictor". It will be convenient to reserve the term *regressor* for a linear combination $c^{\mathrm{T}}x = \sum c_j x_j$ of explanatory variables $x_1, \ldots, x_p$, where the vector c is scaled to unit length, $|c| = 1$.

As far as the first factor is concerned, many methods share the following property: A regressor c is first determined, by some rule, and the predictor $b \propto c$ is then obtained from simple linear regression of y on Xc. Additional factors are constructed from the residuals in accordance with the same rule. Ridge regression is the only method (among those considered here) where this orthogonalization is not applied by definition; our study includes a discussion of the consequences of applying this to RR as well.

We recall that in ridge regression we select the value of a constant δ (the ridge constant) and define the predictor b as

$$b^{\mathrm{RR}}(\delta) = (X^{\mathrm{T}}X + \delta I)^{-1}X^{\mathrm{T}}y.$$

In what follows, we focus on the class of regressors obtained by projecting the vectors $b^{\mathrm{RR}}(\delta)$ onto the unit sphere, i.e. we form regressors

$$c_\delta \propto (X^{\mathrm{T}}X + \delta I)^{-1}X^{\mathrm{T}}y, \quad |c_\delta| = 1. \tag{1}$$

In RR, the ridge constant δ is typically non-negative, but de Jong & Farebrother (1994) pointed out that Sundberg's (1993) relationship could be extended to the interval $-\infty \leqslant \delta < -\lambda_{\max}$, where $\lambda_{\max}$ is the largest eigenvalue of $X^{\mathrm{T}}X$. As δ runs through the two intervals $[0, \infty]$ and $[-\infty, -\lambda_{\max}]$, the point c_δ describes two continuous trajectories on the unit sphere. The limits as $\delta \to \infty$ and $\delta \to -\infty$ are the same, corresponding to first-factor PLS, so we have essentially one trajectory, that extends from OLS at one end ($\delta = 0$), via PLS ($\delta = \pm\infty$) to PCR at the other end ($\delta \to -\lambda_{\max} - 0$) (provided only that data y are not orthogonal to the direction corresponding to the largest eigenvalue, see appendix A). An alternative parametrization of the same trajectory is

$$c_\epsilon \propto \left(X^{\mathrm{T}}X + \frac{\lambda_{\max}}{\epsilon} I\right)^{-1} X^{\mathrm{T}}y, \quad |c_\epsilon| = 1, \tag{2}$$

where ϵ runs through the semi-infinite interval $\epsilon > -1$. In this case, PCR is the limit as $\epsilon \to -1$ and OLS is the limit as $\epsilon \to \infty$. This parametrization avoids the jump from ∞ to $-\infty$ at PLS. A third possibility, also binding the two parts of the trajectory together, is to use the transformed parameter $\delta' = \delta/(\delta + \lambda_{\max})$. This alternative is similar to the parameter γ defined by Stone & Brooks (1990) in that $\delta' = 0$ for OLS, $\delta' = 1$ for PLS and $\delta' \to \infty$ for PCR.

2. Optimality of ridge-type regressors

2.1. A proposition

Near collinearity between explanatory variables means that some linear combinations $c^{\mathrm{T}}x$ vary little between the observations. Under arbitrary scaling of the explanatory variables one cannot exclude that such a linear combination is in fact a major determinant of the response. However, if almost constant linear combinations are allowed as regressors, their OLS regression coefficients will have large uncertainty (even though some of these coefficients might be statistically significant). In particular, these coefficients may be spuriously high and grossly misleading for prediction. One way to protect against this uncertainty is to shrink the OLS predictor in directions where $|Xc|$ is small. This is particularly important if one expects future observations to be different from the training data in some respects. Apart from a constant factor, $|Xc|^2$ equals the sample variance of the linear combination $c^{\mathrm{T}}x$, so $V(c) = |Xc|^2$ is a crucial measure of how useful a regressor c will be, large values of $V(c)$ making less shrinkage needed. Note that first-factor PCR implies that we select c so that $V(c)$ is maximal, without regard to how much or how little correlation with y is obtained by this choice. OLS, on the other hand, selects c for maximum correlation with y, without regard to the size of $V(c)$. Stone & Brooks (1990) defined a continuum of regressors by considering, for each $\gamma \geqslant 0$, the regressor c_γ that maximizes the particular function

$$T_\gamma = R^2(c)V^\gamma(c) \tag{3}$$

of these two measures. From the combined results of Sundberg (1993), de Jong & Farebrother (1994) and Björkström & Sundberg (1996) we know that each such regressor is proportional to a ridge regressor, for some ridge constant, whereas the converse need not hold. Ridge regression is thus in a sense more basic than Stone & Brooks continuum regression. We will now demonstrate that the ridge regressor is even more fundamental than that, by showing that any reasonable maximization criterion which depends only on R and V must yield a regressor proportional to a ridge regressor. The proof is more lucid when formulated in terms of covariance, $K(c) = y^{\mathrm{T}}Xc$, rather than correlation R. We have:

Proposition 2.1
If a regressor c_f is defined according to the rule

$$c_f = \arg\max_{|c|=1} f(K^2(c),\ V(c)), \tag{4}$$

where $f(K^2, V)$ is increasing in K^2 (or $R^2 \propto \mathrm{K}^2/V$) *for constant V, and increasing in V for constant R^2, and if $X^{\mathrm{T}}y$ is not orthogonal to all eigenvectors corresponding to $\lambda_{\max}$, then there exists a number δ such that $c_f \propto (X^{\mathrm{T}}X + \delta I)^{-1}X^{\mathrm{T}}y$, including the limiting cases $\delta \downarrow 0$, $\delta \uparrow \infty$, and $\delta \uparrow -\lambda_{\max}$.*

Remark 1. For any $\gamma > 0$, the Stone & Brooks function (3) evidently satisfies the conditions posed on f.

Remark 2. It seems reasonable that, for a given sample variance $V(c)$, the criterion should favour a regressor with larger covariance, since this improves the correlation coefficient. Thus, f would typically be increasing in K^2 for constant V, as required in the proposition. Likewise, regressors with the same K^2/V have the same R^2, and among them, the one with larger sample variance should be preferred, since it has the best-determined coefficient.

Remark 3. The condition that $X^{\mathrm{T}}y$ not be orthogonal to the eigenspace corresponding to $\lambda_{\max}$ is necessary, but only exceptionally not satisfied. Without it, examples can be constructed where the maximum point is not of the form $c_f \propto (X^{\mathrm{T}}X + \delta I)^{-1}X^{\mathrm{T}}y$. For instance, we may use Stone & Brooks' original criterion $f = R^2 V^{\gamma}$. It turns out that when $\gamma > 1$, there can exist regressors that are not proportional to ridge regressors. We can take $p = 3$, $X^{\mathrm{T}}X =$ diag(1, 0.5, 0.2), $X^{\mathrm{T}}y = (0, 1, 1)^{\mathrm{T}}$, and γ large enough, say $\gamma = 10$. Since the first coordinate of $X^{\mathrm{T}}y$ is zero, all vectors of the form $c \propto (X^{\mathrm{T}}X + \delta I)^{-1}X^{\mathrm{T}}y$ will have their first coordinate $c_1 = 0$. However, the maximum of f on $|c| = 1$ is at $c = (0.910, 0.351, 0.219)^{\mathrm{T}}$.

CR regressors with $\gamma > 1$ typically correspond to ridge regressors with negative ridge parameter, as explained by de Jong & Farebrother (1994). However, these authors tacitly excluded the degenerate case $g_1 = 0$ and therefore overlooked the possibility of situations like in this example.

Remark 4. As a background for the proof, the following picture may be helpful. Every point c on the unit sphere can be represented by a point $(K^2(c), V(c))$ in a two-dimensional diagram as illustrated in Fig. 1. The region Ω is the image of the unit sphere, $|c| = 1$. Figure 1 shows a case where $\lambda_{\min} > 0$, so that $V(c) > 0$ everywhere on the unit sphere. This is not always the case in practice, but it simplifies the picture and has no consequence for the argumentation that follows. In appendix A.1 we comment on the case $\lambda_{\min} = 0$, see also Fig. 5.

Proof of proposition. Since f is increasing in K^2 for constant V, a necessary condition on c_f is that the point $(K^2(c_f), V(c_f))$ fall on the boundary of Ω, on the right hand side.

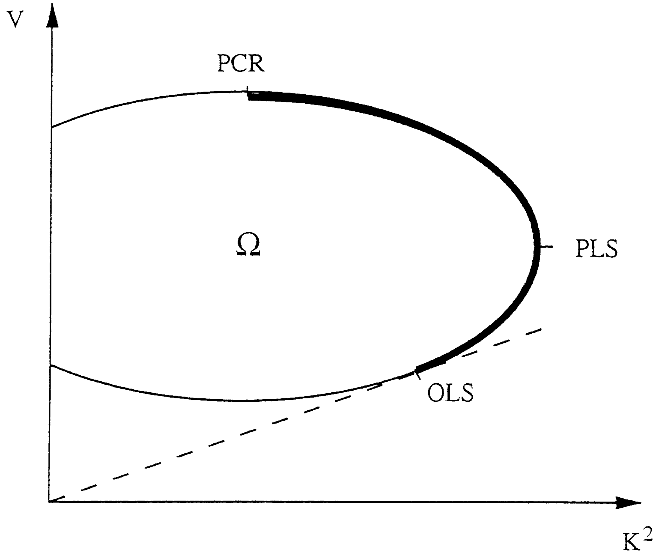

Fig. 1. Illustration of possible (K^2, V)-values, in principle. Each point c on the unit sphere in $\mathbf{R}^p$ is depicted on $(K^2(c), V(c))$. The function $f(K^2, V)$ will have its maximum somewhere on the shaded segment of the perimeter.

Thus, c_f must be a solution to the constrained maximization problem "maximize $K^2(c)$ subject to $|c|^2 = 1$ and $V(c) = v$", for some number v. Since f is increasing in V for constant K^2/V, the relevant number v cannot be smaller than the sample variance of the (minimum-length) ordinary least squares regressor, v_{OLS}. The upper bound for v is $\lambda_{\max}$. Unless v equals this upper bound, it follows from Lagrange's method of multipliers that, at $c = c_f$,

$$\nabla K^2 \in \text{span}[\nabla|c|^2, \nabla V].$$

Thus, there exist two numbers (Lagrangian multipliers) μ_1 and μ_2, such that

$$\nabla K^2 = \mu_1 \nabla|c|^2 + \mu_2 \nabla V,$$

or, since $\nabla K = X^{\text{T}}y$, $\nabla V = 2X^{\text{T}}Xc$, and $\nabla|c|^2 = 2c$,

$$K(c)X^{\text{T}}y = \mu_1 c + \mu_2 X^{\text{T}}Xc.$$

We exclude the case $K = 0$ since it corresponds to the minimum of f, not the maximum. The case $\mu_2 = 0$ yields the well-known limiting case of first-factor PLS, so we concentrate on the case $\mu_2 \neq 0$. We denote the non-zero number K/μ_2 by κ, and μ_1/μ_2 by δ, and get

$$(X^{\text{T}}X + \delta I)c = \kappa X^{\text{T}}y. \tag{5}$$

We show in appendix A.1 that the matrix $X^{\text{T}}X + \delta I$ cannot be singular at the maximum point c_f. Thus it has an inverse and we can write $c_f \propto (X^{\text{T}}X + \delta I)^{-1}X^{\text{T}}y$. In appendix A.2 we consider the limiting case $v = \lambda_{\max}$.

2.2. Other regressor constructions

It follows from proposition 2.1 that if a regression method yields regressors not proportional to ridge regressors, then that method cannot be defined as the maximum of any function of V and R^2 increasing in both arguments. Thus, such a method is not optimal, unless its use can be defended by arguments other than variance and correlation. Among methods of this type are power ridge regression (Hoerl & Kennard, 1975), and the method proposed by Wise & Ricker (1993), also named continuum regression.

3. Least-squares ridge regression

3.1. An alternative function to maximize

We found in section 2 that the set of regressors proportional to ridge regressors (equation 1) contains the f-maximizing regressors for all realistic functions f. A nice feature of the approach taken by Stone & Brooks (1990) is that it specifies a parametric class of functions typically spanning all of this path. Since it maximizes R^2V^γ, the selected regressor is optimal in a well-specified sense, and this optimality may help to understand properties of the predictor. However, as we have here demonstrated, any parametrization of the form

$$c_\gamma = \arg\max_{|c|=1} T_\gamma(R^2(c), V(c)) \tag{6}$$

will define a set of regressors along this particular path, but not necessarily cover the path completely. In particular, the CR parametrization used by Stone & Brooks can make jumps, as demonstrated by Björkström & Sundberg (1996). It turns out that there exist parametrizations free from jumps. One possibility is given by the following criterion, which, to our knowledge, was first stated by Frank & Friedman (1993):

$$c_\delta = \arg\max_{|c|=1} \frac{R^2(c)V(c)}{|V(c)+\delta|}. \tag{7}$$

Frank & Friedman showed that formula (7) for $\delta \geqslant 0$ defines a regressor c_δ that is proportional to the ridge regressor $b^{\mathrm{RR}}(\delta) = (X^{\mathrm{T}}X + \delta I)^{-1}X^{\mathrm{T}}y$ with the same δ. It is evident that the function $b^{\mathrm{RR}}(\delta)$ is free from discontinuities. Apart from notational differences, our formula (7) differs from definition (20) in Frank & Friedman (1993) only so as to allow $\delta < -\lambda_{\max}$ (cf. appendix B).

By analogy with how Stone & Brooks (1990) defined continuum regression based on the one-parametric class of regressors defined by equations (3) and (6), for $0 \leqslant \gamma \leqslant \infty$, we now define what may be called least-squares ridge regression (LSRR): We consider the class of regressors defined by (1) and the parameter δ. The corresponding predictors are

$$b^{\mathrm{LSRR}}(\delta) = \beta c_\delta, \tag{8}$$

where the scalar factor β is

$$\beta = y^{\mathrm{T}}Xc_\delta/|Xc_\delta|^2. \tag{9}$$

Stone & Brooks consider an arbitrary number ω of factors, and use cross-validation to determine the optimal pair (γ, ω). A completely analogous procedure is possible with LSRR, yielding an optimal pair (δ, ω).

3.2. Correlation and shrinkage properties

Methods for treating near-collinearity typically yield predictors whose lengths are smaller than that of the OLS predictor. This "shrinkage effect" is closely coupled with the variance reduction and bias of the predictor. Frank & Friedman (1993) compared RR, PCR and PLS in this respect, in a number of situations with varying degree of collinearity. Goutis (1996) and De Jong (1995), using different techniques, have shown that PLS regression always shrinks, as long as the number of factors is smaller than the number of explanatory variables, p. De Jong (1993) has also demonstrated that the coefficient of determination obtained by PCR with a given number of factors, cannot be larger than obtained by PLS with the same number of factors. By considering the class of regressors (1) and limiting ourselves to one factor ($\omega = 1$), we can generalize the results of these authors, by proving the following two propositions.

Proposition 3.1
$R^2(c_\delta)$ decreases along the path from OLS via PLS to PCR.

Proposition 3.2
The length of the predictor coefficient vector, $|b^{\mathrm{LSRR}}(\delta)|$ decreases along the path from OLS via PLS to PCR.

In other words, proposition 3.1 states that the fit between y and $\hat{y}$ is gradually reduced as one moves away from the OLS predictor, and proposition 3.2 states that the LSRR estimator $b^{\mathrm{LSRR}}(\delta)$ is a regularized estimator (except for OLS), shrinking in length as one moves from OLS towards PCR. However, note that $b^{\mathrm{LSRR}}(\delta)$ need not shrink in all its coordinates.

We want to point out that we prove propositions 3.1 and 3.2 only for first-factor regression. In section 4, we show by counterexample that the corresponding result for an arbitrary number of factors does not hold for proposition 3.2.

The proofs of propositions 3.1 and 3.2 are given in appendix B.

4. Applications to data

We have tested our procedure on a limited number of data sets. Our main purposes at present are to compare our results to those obtained using original CR and to illustrate the conclusions drawn in propositions 3.1 and 3.2. An evaluation of the LSRR method, based on a more extensive selection of data sets, is being prepared and will be reported elsewhere. Factors beyond the first one have been constructed keeping δ constant (although other possibilities also exist, see conclusions).

Our main data set is taken from Fearn (1983). These data have been used for examplification by many authors. We follow Stone & Brooks (1990) and consider the subset consisting of the first $n = 12$ items of Fearn's table 1. Rather than using the explanatory variables as they stand ($L_1, \ldots, L_6$) we have, like Hoerl *et al.* (1985) and Stone & Brooks (1990), transformed by defining $x_6 = (L_1 + \cdots + L_6)/6$ and $x_j = L_j - x_6$ for $j = 1, \ldots, 5$. All explanatory variables have been scaled to unit variance.

4.1. Cross-validation index as criterion for predictor choice

We have constructed predictors for Fearn's data with $n = 12$ by applying our procedure for several different combinations of number of factors, ω, and ridge constant, δ. For the evaluation, we have used the same leave-one-out cross-validation index I as Stone & Brooks (1990).

Figure 2 shows $I = I_\delta$ as a function of δ for $1 \leqslant \omega \leqslant 5$. On the horizontal axis, we have plotted $\delta/(\delta + \lambda_{\max})$ rather than δ itself. This admits presentation of the two intervals $0 \leqslant \delta \leqslant \infty$ and $-\infty \leqslant \delta \leqslant -\lambda_{\max}$ in the same figure, the former interval corresponding to $0 \leqslant \delta/(\delta + \lambda_{\max}) \leqslant 1$, and the latter to $\delta/(\delta + \lambda_{\max}) \geqslant 1$. As δ runs from OLS to PCR, our equation (1) defines regressors that could also have been defined by maximizing Stone & Brooks' criterion (3) for some $0 \leqslant \gamma \leqslant \infty$. However, there is no one-to-one correspondence between their parameter γ and our δ when $\omega \geqslant 2$, since holding γ constant to the next factor is not equivalent to holding δ constant. As far as leave-one-out cross-validation is concerned, there is no exact correspondence even for $\omega = 1$, since the relation between γ and δ will change with the observation that is left out. Stone & Brooks' parameter γ is roughly comparable to our transformed parameter $\delta/(\delta + \lambda_{\max})$ insofar as both are 0 for OLS, 1 for PLS and ∞ for PCR.

It is of interest to compare our Fig. 2 to fig. 3 of Stone & Brooks (1990). Their value $\alpha = 0.5$ is PLS regression, equivalent to $\delta = \infty\,(\delta/(\delta + \lambda_{\max}) = 1$ in our parametrization). The two figures are similar. For the first factor ($\omega = 1$), I increases in a short range of parameter values (approximately $\delta < 10^{-1}$), and then rapidly drops to uninteresting levels. In contrast, for $\omega \geqslant 2$, the I-value stays above 0.9 for the whole range between OLS and PLS. From the point of view of cross-validation, our optimal choice of parameters is $\delta = 2.10$ and $\omega = 3$, which gives $I_\delta = 0.9592$. This is not much different from $I_\alpha = 0.960$, the optimal value obtained by Stone & Brooks with $\omega = 2$. Table 1 shows the result.

When not standardizing the explanatory variables for sample variance, Stone & Brooks obtain a solution which is different in that the coefficient of L_1 is about four times larger and of

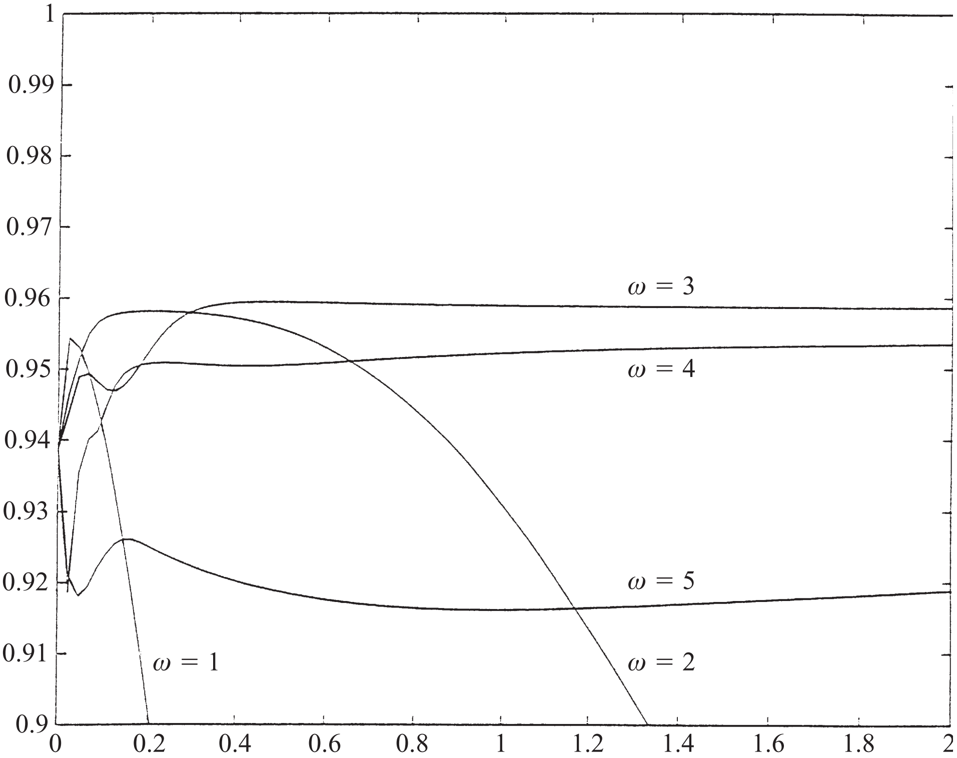

Fig. 2. Leave-one-out cross-validation index I_δ plotted vs $\delta/(\delta + \lambda_{\max})$, for LSRR with varying number of factors (ω). The figure is based on the first 12 observations of Fearn's (1983) data. The explanatory variables are standardized, which gives $\lambda_{\max} = 3.19$.

opposite sign. All other coefficients change relatively less than so. With our method, the effect of not standardizing is also most pronounced for the coefficient of L_1, which becomes roughly seven times larger, but keeps its sign. In neither case is L_1 among the most important predictors. In terms of the untransformed variables x_j, the effect of not standardizing is in our case largest for x_5. This differs from the result by Stone & Brooks, where x_1 is affected most.

Lines 3 and 6 of Table 1 show the best solution obtained under the additional constraint $\omega = 1$, i.e. using only the first factor. We note that this extra constraint does not reduce I_δ much (from 0.9592 to 0.9544). The effect on the predictor is also modest. The decision to stop or proceed after the first step seems to affect the predictor about as much as does the decision whether to standardize or not.

4.2. Illustrations of propositions 3.1 and 3.2

Coefficient of determination, R^2. It is of interest to see how much of the variation in the y data is explained by the regression models when the number of factors and the parameter δ are varied. Figure 3 shows the picture for Fearn's data with $n = 12$. In accordance with proposition 3.1, R^2 decreases with δ for the first factor ($\omega = 1$). Note that with $\omega \geqslant 3$, at least 0.975 of the variation is explained, regardless of δ. The figure confirms (of course) de Jong's result "PLS fits closer than PCR", but we leave open the question whether R^2 for $\omega \geqslant 2$ can possibly increase locally with δ.

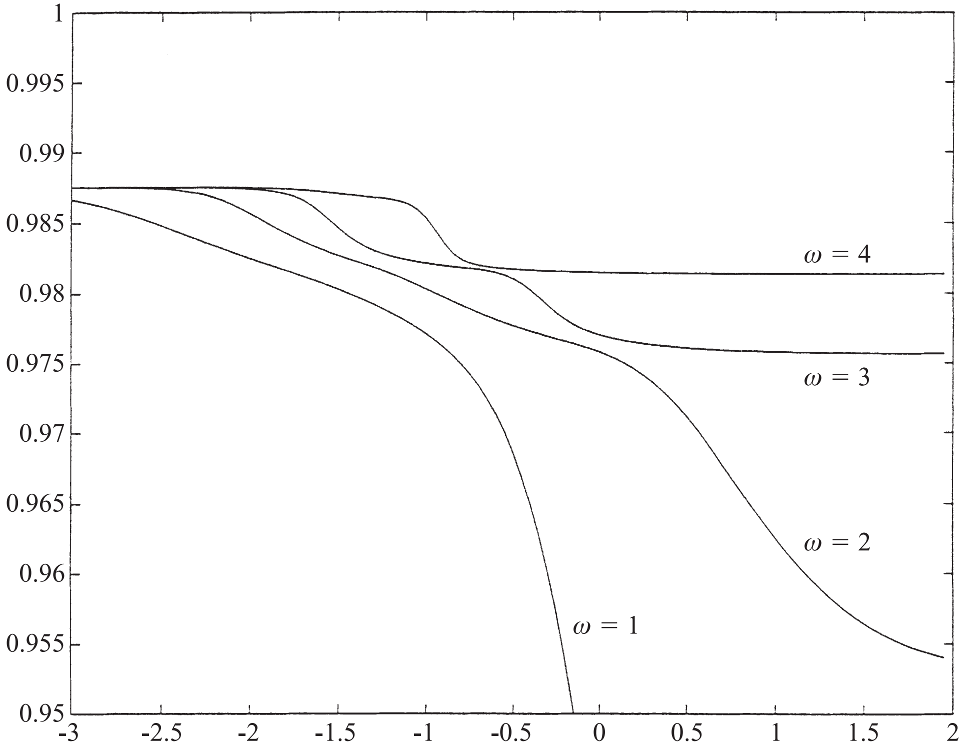

Fig. 3. Coefficient of determination, R^2, as a function of the parameter, for LSRR with $1 \leqslant \omega \leqslant 4$. The abscissa shows $\log_{10}(\delta)$.

Table 1. *Coefficients in the optimal predictor as defined by leave-one-out cross-validation. The results with parameter γ were taken from Stone & Brooks. The table is based on the first 12 observations of Fearn's (1983) data*

		Standardized data						
Parameter values	$\delta/(\delta + \lambda_{\max})$	$I_{\max}$	L_1	L_2	L_3	L_4	L_5	L_6
$\omega = 2$, $\gamma = 0.54$		0.960	−0.02	0.09	0.16	−0.21	0.01	−0.01
$\omega = 3$, $\delta = 2.10$	0.40	0.9592	0.000	0.074	0.140	−0.181	0.004	−0.009
$\omega = 1$, $\delta = 0.086$	0.026	0.9544	0.014	0.082	0.159	−0.226	0.005	−0.008
		Unstandardized data						
Parameter values	$\delta/(\delta + \lambda_{\max})$	$I_{\max}$	L_1	L_2	L_3	L_4	L_5	L_6
$\omega = 4$, $\gamma = 0.43$		0.958	0.09	0.08	0.13	−0.27	0.01	−0.01
$\omega = 4$, $\delta = 84$	0.024	0.9496	0.002	0.124	0.127	−0.204	−0.017	0.001
$\omega = 1$, $\delta = 23$	0.007	0.9473	−0.006	0.123	0.126	−0.210	0.005	−0.002

Degree of shrinkage. Figure 4 shows the length of the predictor b^{LSRR} for a range of δ values. For $\omega = 1$, it is monotone decreasing, in accordance with proposition 3.2. The graphs for $\omega = 2$ and $\omega = 3$ are interesting in that the length increases with δ near OLS. We thus see that proposition 3.2 cannot be generalized to hold for more than the first-factor regressor.

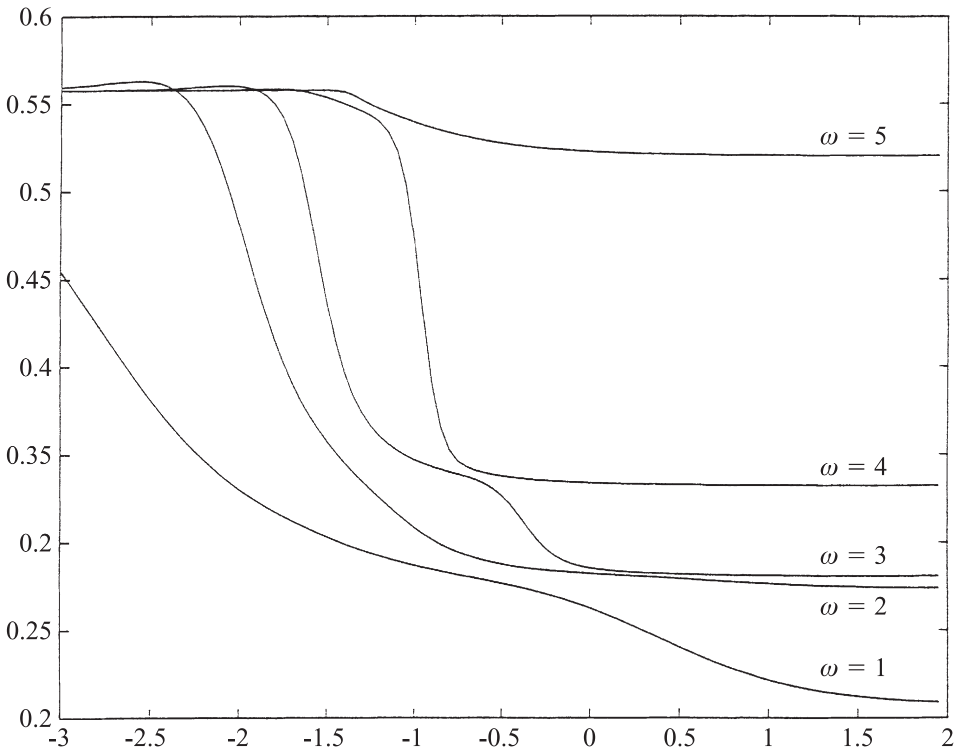

Fig. 4. Illustration of shrinkage effect. The length of the predictor, $|b^{\mathrm{LSRR}}|$, is plotted vs δ for different number of factors ω. The figure is based on the first 12 observations of Fearn's (1983) data. For $\omega = 1$, the length is a decreasing function of δ (proposition 3.2), but this is not the case for $\omega > 1$.

5. Conclusions

Least squares ridge regression, LSRR, is a regression method that improves on ridge regression, RR, in the following respects.

(1) Avoids unwanted shrinkage. The shrinkage in standard RR can be regarded as the joint effect of two separate principles: to counterbalance the inflating effect of near-collinearities and to downscale any unbiased estimator in order to reduce the MSE. For those who do not admit the latter principle as a general one, LSRR offers a way to concentrate on the near-collinearity effects. Note for example that LSRR does not shrink in the case of a single x-variable or in orthogonal designs, whereas RR does so. Note also that Sundberg (1993) empirically found LSRR equally efficient to RR but more robust to the choice of ridge parameter.

(2) Orthogonal residuals. LSRR leaves residuals that are orthogonal to the fitted values. This facilitates the interpretation of factors beyond the first one. Another consequence is that the fitted values $\hat{y} = Xb^{\mathrm{LSRR}}(\delta)$ are as close as possible, in the chosen metric, to the observed y, given the direction of $\hat{y}$.

LSSR shares these two properties with standard continuum regression, CR. In addition, LSRR improves on CR in the following respects.

(3) No discontinuity. With LSRR, the predictor always varies continuously with the method parameter, for a fixed number of factors.

(4) Opens up for alternative parametrizations. When LSRR is used iteratively, there is no reason why the method parameter δ must have the same value in each step. Many other principles are conceivable. For example PCR amounts to holding δ equal to minus the largest eigenvalue of the current $X^{\mathrm{T}}X$, although this number will gradually decrease with each step, as the residual matrix X is updated.

(5) Speed of computations. The computer time required for application of equation (1) with given δ is substantially less than for finding the maximum of R^2V^γ for a given γ. The maximization (7) is explicit and involves much less computational effort than does CR.

References

Björkström, A. & Sundberg, R. (1996). Continuum regression is not always continuous. *J. Roy. Statist. Soc. Ser. B* **58**,703–710.

Brown, P. J. (1993). *Measurement, regression, and calibration.* Oxford University Press, Oxford.

de Jong, S. (1993). PLS fits closer than PCR. *J. Chemometrics* **7**, 551–557.

de Jong, S. (1995). PLS shrinks. *J. Chemometrics* **9**, 323–326.

de Jong, S. & Farebrother, R. W. (1994). Extending the relationship between ridge regression and continuum regression. *Chemometrics Intelligent Lab. Systems* **25**, 179–181.

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

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

Goutis, C. (1996). Partial least squares algorithm yields shrinkage estimators. *Ann. Statist.* **24**, 816–824.

Hoerl, A. E. & Kennard, R. W. (1975). A note on a power generalization of ridge regression. *Technometrics* **17**, 269.

Hoerl, A. E., Kennard, R. W. & Hoerl, R. W. (1985). Practical use of ridge regression: a challenge met. *J. Appl. Statist.* **34**, 114–120.

Stone, M. & 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. Roy. Statist. Soc. Ser. B* **52**, 237–269; corrigendum (1992) **54**, 906–907.

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

Sundberg, R. (1999). Multivariate calibration—direct and indirect regression methodology (with discussion). *Scand. J. Statist.*, to appear.

Wise, B. M. & Ricker, L. N. (1993). Identification of finite impulse response models with continuum regression. *J. Chemometrics* **7**, 1–14.

Received July 1996, in final form March 1998

Anders Björkström, Mathematical Statistics, Stockholm University, S-106 91 Stockholm, Sweden.

Appendix A. The remainder of the proof of proposition 2.1

It was shown in the main text that if a regressor is defined as the maximum of a function $f(K^2(c), V(c))$ on the unit sphere in p-dimensional space, this regressor must be a solution to the problem

$$\text{``maximize } K^2(c) \text{ subject to } |c|^2 = 1 \text{ and } V(c) = v\text{''},$$

for some number v. We showed that if v is in the interval $[v_{\mathrm{OLS}}, \lambda_{\max})$, then an equation like (5) holds. It remains to be shown that $X^{\mathrm{T}}X + \delta I$ is non-singular, and to prove that the regressor is ridge-proportional also when $v = \lambda_{\max}$.

A.1. Why the matrix in equation (5) is not singular

Since the eigenvectors of $X^{\mathrm{T}}X$ constitute an orthonormal basis (the canonical basis), we can without loss of generality assume that $X^{\mathrm{T}}X$ is diagonal, $X^{\mathrm{T}}X = \Lambda = \mathrm{diag}(\lambda_1 \ldots, \lambda_p)$, where $\lambda_{\max} = \lambda_1 \geqslant \cdots \geqslant \lambda_p = \lambda_{\min} \geqslant 0$, and $X^{\mathrm{T}}y$ equals a vector g such that $g_j \geqslant 0$ for all j. By the assumptions in the proposition, we have $g_1 > 0$. In canonical form, the Lagrangian equation (5) reads

$$(\Lambda + \delta I)c = \kappa g, \tag{10}$$

with $\kappa \neq 0$. We need to ascertain that the particular δ-value that holds at the global maximum c_f is not equal to $-\lambda_j$ for any j. Since all $g_j \geqslant 0$, all coordinates of c_f are non-negative, too. (Otherwise we could change the signs of the negative ones and increase $K^2 = (\Sigma c_j g_j)^2$ without violating either of the two constraints.) Thus, if $\delta = -\lambda_j$ for some λ_j, this must be either for λ_1 or λ_p, otherwise equation (10) would imply that c_f had both positive and negative coordinates. Furthermore, by the assumptions we have $g_1 \neq 0$, so the right hand side of equation (10) is non-zero for $j = 1$, which rules out the possibility $\delta = -\lambda_1$ also.

Comment. The example mentioned in remark 3 illustrates what may happen when $g_1 = 0$. The essential ingredients of that example are $g_1 = 0$ and $v > \lambda_2$, the second largest eigenvalue. Under these circumstances, it is impossible to satisfy $V(c) = v$ with $c_1 = 0$. Equation (10) still holds, with $\delta = -\lambda_1$, but the solution cannot be written $c_f \propto (\Lambda + \delta I)^{-1} g$.

To rule out the one remaining possibility, $\delta = -\lambda_p$, suppose that c_0 is a solution to equation (10) for $\delta = -\lambda_p$. For such a solution to exist, it is necessary that $g_p = 0$. Note that generally, $V(c) = \Sigma + \lambda_j c_j{}^2$ can be regarded as a weighted average of the eigenvalues λ_j, the weights being $c_j{}^2$. For OLS we have $c_j{}^2 \propto (g_j/\lambda_j)^2$, and for the point c_0 we have $c_j{}^2 \propto (g_j/(\lambda_j - \lambda_p))^2$. These two expressions differ only in the denominator. By subtracting the same number λ_p from all the eigenvalues we increase the weights proportionately more for small λ_j than for big ones. Thus, if $\lambda_p > 0$, $V(c_0) < V(c_{\mathrm{OLS}})$. As already demonstrated, a point c_0 where this strict inequality holds cannot be global maximum for the function f. If $\lambda_p = 0$, the inequality is not strict, and we cannot exclude that c_0 might be the global maximum. However, when $\lambda_p = 0$, the region Ω extends down to the origin, as indicated in Fig. 5. All points on the correlation-maximizing straight line between the origin and the point marked OLS correspond to least-squares regressors: $c_j \propto g_j/\lambda_j$ for $j \leqslant r$, c_j arbitrary for $j > r$, and all are solutions to (10) with $\delta = -\lambda_p (= 0)$. The point marked OLS corresponds to the least-squares regressor with largest variance, $c_j = 0$ for $j > r$. This is the only point on the line that can correspond to the global maximum c_f, because of the assumptions on f, and it corresponds to the limiting case $\delta \downarrow 0$, as formulated in the proposition.

In conclusion, disregarding the limiting case just discussed, the matrix in equation (5) is not singular at the maximum point c_f.

A.2. The limiting case $v = \lambda_{max}$

We finish the proof of proposition 2.1 by considering the case that the maximum of f occurs at a point c_f which is represented by the top of the region Ω, i.e. when $V(c)$ is maximal, $V(c) = \lambda_{\max} = \lambda_1$. Unless the eigenvalue λ_1 has multiplicity $\geqslant 2$, there is a unique point $c = v_1$ of the unit sphere that satisfies $V(c) = \lambda_1$. Lagranges method of multipliers is not applicable then, but provided $g_1 > 0$, v_1 is the limit of $c \propto (\Lambda + \delta I)^{-1} g$ as $\delta \uparrow -\lambda_1$.

If the eigenvalue λ_1 has multiplicity $m_1 \geqslant 2$, a continuum of points c with $|c| = 1$ will satisfy

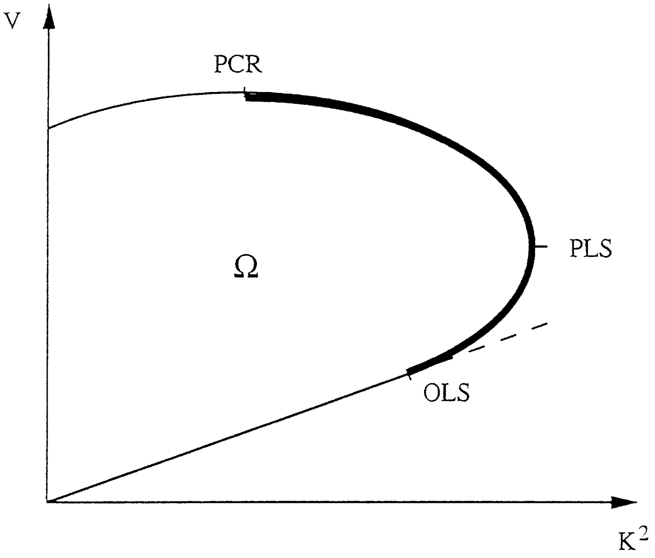

Fig. 5. Image of the unit sphere in a case where $\lambda_{\min} = 0$.

the constraint $V(c) = \lambda_1$. It is easily realized (using Cauchy–Schwarz' inequality, for example) that K^2 is maximized by taking $c_{\mathrm{j}} \propto g_j$ for $j \leqslant m_1$, and $c_j = 0$ for $j > m_1$. This is also exactly the point towards which $c \propto (\Lambda + \delta I)^{-1} g$ converges as $\delta \uparrow -\lambda_1$, provided $g_j > 0$ for all $j \leqslant m_1$. (This can always be provided since the vectors $v_1, \ldots, v_{m_1}$ are not unique.) Thus, when $v = \lambda_1$, the regressor c_f is proportional to the limit of a ridge regressor, as asserted in proposition 2.1. The proof of the proposition is thereby complete.

Appendix B. Proof of propositions 3.1 and 3.2

The proofs of propositions 3.1 and 3.2 are simplified if we replace the parameter δ by a new parameter $\varepsilon = \lambda_{\max}/\delta$. The two disjoint intervals for δ correspond to one (connected) interval for ε: $-1 < \varepsilon < \infty$. Any function increasing in δ will be decreasing in ε, and vice versa. Formula (7) is replaced by

$$c_\varepsilon = \arg\max_{|c|=1} T(c, \varepsilon) \tag{11}$$

where

$$T(c, \varepsilon) = \frac{R^2(c)|Xc|^2}{\lambda_{\max} + \varepsilon|Xc|^2}. \tag{12}$$

Note that since $\varepsilon > -1$ and $|Xc|^2 \leqslant \lambda_{\max}$ on $|c| = 1$, $\mathrm{T}(c, \varepsilon) \geqslant 0$ for all ε and all c on the unit sphere.

Proof of proposition 3.1. We must show that $R^2(c_\varepsilon)$ is an increasing function of ε. Rearranging in (12) we get

$$\frac{\lambda_{\max}}{|Xc|^2} = \frac{R^2(c)}{T(c, \varepsilon)} - \varepsilon.$$

Here, the left hand side does not contain ε. Therefore, for any vector c and any two numbers ε_1 and ε_2 it must hold that

$$\frac{R^2(c)}{T(c,\,\varepsilon_1)} - \varepsilon_1 = \frac{R^2(c)}{T(c,\,\varepsilon_2)} - \varepsilon_2,$$

which yields

$$\frac{R^2(c)}{\varepsilon_2 - \varepsilon_1} = \left[\frac{1}{T(c,\,\varepsilon_2)} - \frac{1}{T(c,\,\varepsilon_1)}\right]^{-1}. \tag{13}$$

We now take $\varepsilon_1 < \varepsilon_2$ and apply equation (13) with corresponding optimal c_1 and c_2. From (11) and (12) we obtain the following sequence of inequalities:

$$T(c_1,\,\varepsilon_2) \leqslant T(c_2,\,\varepsilon_2) \leqslant T(c_2,\,\varepsilon_1) \leqslant T(c_1,\,\varepsilon_1), \tag{14}$$

where the middle inequality holds because when c is constant, the T of (12) is decreasing in ε for all $\varepsilon > -1$. It follows from (13) and (14) that $R^2(c_1) \leqslant R^2(c_2)$, proving that $R^2(c_\varepsilon)$ is an increasing function of ε. This is just the statement of proposition 3.1.

Proof of proposition 3.2. To prove that $|b^{\mathrm{LSRR}}(\varepsilon)|$ is increasing, we note from (8) and (9) that, since $|c_\varepsilon| = 1$,

$$|b^{\mathrm{LSRR}}(\varepsilon)|^2 = \frac{|y^{\mathrm{T}} X c_\varepsilon|^2}{|X c_\varepsilon|^4} = \frac{|y|^2 R^2(c_\varepsilon)}{|X c_\varepsilon|^2}.$$

We have proven that $R^2(c_\varepsilon)$ increases, so proposition 3.2 follows if we show that $1/|Xc_\varepsilon|^2$ also increases. With c_1 and c_2 as above, we need to prove that $1/|Xc_1| \leqslant 1/|Xc_2|$ for all $\varepsilon_1 < \varepsilon_2$.

By rearranging terms in (12) once more we get

$$R^2(c) = T(c,\,\varepsilon)\left(\varepsilon + \frac{\lambda_{\max}}{|Xc|^2}\right)$$

where the left hand side contains no ε. We conclude that for any triple $(c,\,\varepsilon_1,\,\varepsilon_2)$ we have

$$T(c,\,\varepsilon_1)\left(\varepsilon_1 + \frac{\lambda_{\max}}{|Xc|^2}\right) = T(c,\,\varepsilon_2)\left(\varepsilon_2 + \frac{\lambda_{\max}}{|Xc|^2}\right)$$

from which we obtain

$$\frac{\lambda_{\max}}{|Xc|^2} = \frac{\varepsilon_2 T(c,\,\varepsilon_2) - \varepsilon_1 T(c,\,\varepsilon_1)}{T(c,\,\varepsilon_1) - T(c,\,\varepsilon_2)} = \frac{(\varepsilon_2 - \varepsilon_1) T(c,\,\varepsilon_2)}{T(c,\,\varepsilon_1) - T(c,\,\varepsilon_2)} - \varepsilon_1$$

The inequalities (14) imply that both $T(c,\,\varepsilon_2)$ and $1/(T(c,\,\varepsilon_1) - T(c,\,\varepsilon_2))$ are greater for $c = c_2$ than for $c = c_1$, and hence, since $\varepsilon_2 - \varepsilon_1 > 0$, that $\lambda_{\max}/|Xc_1|^2 \leqslant \lambda_{\max}/|Xc_2|^2$. As remarked above, this proves proposition 3.2.