Mathematical Statistics
Stockholm University

CONTINUUM REGRESSION
Article for 2nd ed. of Encyclopedia of Statistical Sciences

Rolf Sundberg

Research Report 2002:4

ISSN 1650-0377

Postal address:
Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

Internet:
http://www.matematik.su.se/matstat

Mathematical Statistics
Stockholm University
Research Report **2002:4**,
http://www.matematik.su.se/matstat

CONTINUUM REGRESSION
Article for 2nd ed. of Encyclopedia of Statistical Sciences

Rolf Sundberg*

May 2002

Abstract

When, in a multiple regression, regressors are near-collinear, so called regularized or shrinkage regression methods can be highly preferable to ordinary least squares, by trading bias for variance. Continuum regression, introduced by Stone and Brooks in 1990, ties together several more classical regularized regression methods, such as principal components regression, partial least squares regression, and ridge regression.

Keywords: Collinearity, Cross-validation, Multicollinearity, Partial least squares regression, Principal component regression, Ridge regression, Regularized regression, Shrinkage.

AMS Subject Classification: 62J05, 62J07

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: rolfs@matematik.su.se. Financial support from The Swedish Research Council is gratefully acknowledged.

Collinearity and regularized regression

Continuum regression is a regularized regression estimation method, and being so, it is particularly intended for dealing with the collinearity problem. Collinearity, or multicollinearity*, means that there are approximate (or possibly even exact) linear relationships between the regressors (predictors, x-variables). Let the regression to be fitted be $\mathbf{y} = \mathbf{X}\beta$, in centered x- and y-variables. Collinearity implies that the design matrix $\mathbf{X}$ is close to being of less than full rank. Consequently, $\mathbf{X}'\mathbf{X}$ is (near-)singular, and ordinary least squares* (OLS) regression coefficients have highly inflated variances, since $\mathrm{Var}(b_{OLS}) \propto (\mathbf{X}'\mathbf{X})^{-1}$, and they are likely to be quite unstable or possibly even non-unique. The regression coefficients cannot be interpreted individually. In this sense identification of a true regression will be (nearly) impossible. However, satisfactory prediction may still be quite feasible, if the new items are like the ones used in the calibration. Below we shall mainly have in mind the construction of a linear *predictor* for future y-values. Therefore, cross-validation* and related validation techniques will play an important role.

Exact collinearities are evidently unavoidable if there are more x-variables than observations. With instruments capable of registering a large set of variables, for example a spectrum of hundreds of wavelengths, this is a frequent situation in practice. However, even if there are sufficiently many observations to guarantee a full rank design matrix, most variability in x is likely to concentrate in a relatively small-dimensional space. This non-rigorously defined dimension is the "latent dimension" or (in the chemometrics literature) the "chemical rank" of the system under measurement.

Well-known examples of regularization methods are principal components regression* (PCR) and partial least squares* (or projection to latent structures) regression (PLSR), which both attempt to find and span such a latent space, whereas ridge regression* (RR) performs quite different by shrinking more or less in each direction. The concept of continuum regression (CR), as introduced by Stone & Brooks [9], embraces OLS, PLSR as well as PCR, thus before we define CR, we shall briefly describe these three special cases from a point of view which will naturally lead to CR.

OLS, PCR and PLSR

OLS may be characterized as *maximizing correlation*. The multiple correlation coefficient R is the maximum over all direction vectors $\mathbf{c}$ of the correlation $R(\mathbf{t}, \mathbf{y})$ between $\mathbf{y}$ and a regressor $\mathbf{t} = \mathbf{X}\mathbf{c}$. The desired direction $\mathbf{c}$ of course satisfies $\mathbf{c} \propto b_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. Regression of $\mathbf{y}$ on $\mathbf{t}$, by simple one-dimensional least squares, yields the OLS multiple regression of $\mathbf{y}$ on $\mathbf{X}$.

The PCR and PLSR methods depend on other maximization criteria for construction of regressors from $\mathbf{X}$. In a standard PCR the first regressor is the first principal component (PC), formed by letting $\mathbf{c}_1$ be the direction of highest variability in x, that maximizes the *variance* $\mathrm{Var}(\mathbf{t}_1)$ for $\mathbf{t}_1 = \mathbf{X}\mathbf{c}_1$, $|\mathbf{c}_1| = 1$. Successive regressors $\mathbf{t}_2 = \mathbf{X}\mathbf{c}_2$ etc. are obtained by repeating the procedure on the residuals from the preceding step, and the new PCs will be uncorrelated with the preceding ones. The number of PCs to be used jointly as regressors will typically be optimized by cross-validation, or by using a separate test set. PLSR selects regressors (PLS-factors) $\mathbf{t}_1$, $\mathbf{t}_2$ etc. by maximizing the *covariance* $\mathrm{cov}(\mathbf{t}, \mathbf{y})$, rather than the variance. Stone and Brooks [9] were among the first to realize this fact. In other respects, however, PLSR and PCR are analogous.

Typically, PLSR requires fewer PLS-factors than PCR needs PCs; this is because PCs need not necessarily be correlated with y, whereas all PLS-factors must be. The reason that PCR may provide a useful regression equation at all, is that it avoids directions $\mathbf{c}$ with small variation, and those are the ones which may cause the collinearity problems. PLSR may be regarded as a compromise between OLS and PCR, $\mathrm{cov}^2(\mathbf{t}, \mathbf{y})$ being proportional to the product $R^2(\mathbf{t}, \mathbf{y})\,\mathrm{Var}(\mathbf{t})$.

Continuum regression

In continuum regression the construction rule for new regressors encompasses the three constructions discussed above, which correspond to special values of a control parameter γ selected within a continuum of possible values. The construction rule in [9] can be expressed as follows: Maximize the function

$$g_\gamma(R^2(\mathbf{t}, \mathbf{y}), \mathrm{Var}(\mathbf{t})) = R^2(\mathbf{t}, \mathbf{y})\,\mathrm{Var}(\mathbf{t})^\gamma, \tag{1}$$

over directions $\mathbf{c}$, with $|\mathbf{c}| = 1$, where $\mathbf{t} = \mathbf{X}\mathbf{c}$ and $\gamma \geq 0$. For $\gamma = 0$ we have OLS, $\gamma = 1$ yields the PLSR criterion, and as $\gamma \to \infty$ we obtain PCR in the limit. As is the case with PLSR and PCR, the construction rule is applied first on the centered (or centered and scaled) x-data, and next successively on the residuals from the regression in each step, such that the regressors $\mathbf{t}_1$, $\mathbf{t}_2$, etc. are uncorrelated. For each set of regressors, $\mathbf{y}$ is then regressed on it, using the least squares method.

It is assumed in [9] that we select the optimal control parameter values, γ and the number of regressors, using cross-validation, but in practice we may, of course, weigh in the aspects of parsimony and simplicity. As a cross-validation criterion we may use anyone of the many equivalent criteria commonly used, such as: PRESS, MSEP, RMSEP, or the cross-validation index, analogous to R^2 and sometimes denoted Q^2.

Continuum regression is *scale-dependent*, a property it shares with other shrinkage regressisions, such as PLSR, PCR, and RR. As compared with OLS, CR penalizes regressors with small variance, and the variance is not invariant under non-orthogonal variable transformations. The user must decide what is a reasonable and desirable metric in x-space. Often the x-variables are individually "autoscaled" to unit variances, but this cannot be always recommended, since it might simply blow up the noise in noninformative x-variables.

Continuum regression and ridge regression

Examining more closely the construction rule (1), one observes that it yields a regressor $\mathbf{t} = \mathbf{X}\mathbf{c}$ which is proportional to the *ridge regression** of $\mathbf{y}$ on $\mathbf{x}$, for some ridge constant δ dependant on γ. Hence, CR is closely related to RR. Specifically, we may interpret first factor continuum regression, CR(1), simply as an upscaled ridge regression, with $1/(1-\gamma)$ as the scale factor ([10], [5]). A further insight into CR is provided in [2]: Not just for the special form (1), but for *any* function $g(R^2(\mathbf{t}, \mathbf{y}), \mathrm{Var}(\mathbf{t}))$ of correlation and variance, satisfying the natural condition of being increasing in each one of its two arguments, the maximizing regressor $\mathbf{t} = \mathbf{X}\mathbf{c}$ is of the ridge type, namely

$$\mathbf{c} \propto b_{RR} = (\mathbf{X}'\mathbf{X} + \delta\mathbf{I})^{-1}\,\mathbf{X}'y \tag{2}$$

for some ridge constant δ that depends on the particular function g. The OLS is obtained for $\delta = 0$, and we arrive at PLSR as $\delta \to \pm\infty$, while PCR appears in the limit as δ approaches the minus of the largest eigenvalue of $\mathbf{X}'\mathbf{X}$ (from below).

The RR always shrinks, even if there is only a single x-variable, or the x-variables are orthonormal. On the other hand, CR(1) shrinks only to compensate for collinearity. Hence, for statisticians which are not adhering to the principle of "always shrink", this motivates a modification of RR by the scalar compensation factor $1/(1-\gamma)$ built into CR(1), simply by using $\mathbf{t}_1 = \mathbf{X}b_{RR}$ as a single regressor in LS regression, instead of b_{RR} as an estimator. This modified RR was called *least squares ridge regression* (LSRR) in [2]. There are only two minor differences between LSRR and the CR(1) as described in [9]. The first is that occasionally CR(1) can be found to jump over an interval of the δ-values [1]. Secondly, the cross-validation procedures turn out to be slightly different, because different parameters, δ and γ respectively, are kept fixed. The latter effect is more pronounced if LSRR is used as an alternative to the original CR for construction of additional regressors.

Example

Figure 1 shows cross-validation leave-one-out RMSEP curves for varying control parameters, when RR, LSRR, and CR(k) for k factors are used on the cement heat evolution data, provided already in Hald's classical book of 1952 [7], and used later on in many textbooks, and also in [9]. Characteristic for these data is the fact that the four x-variables represent the composition of the cement and sum up to nearly 100 per cent. The OLS coefficients are large and individually nonsignificant. In this illustration, the original x-variables have been autoscaled, as it is also done in [9]. The continuum parameter which is kept fixed in CR(k) is the ridge constant δ of (2), and not the γ of criterion (1). This implies that LSRR and CR(1) are identical. With four factors in CR, we get the OLS, shown as a horizontal line. The main feature of Fig. 1 is the similarity of the minimum RMSEP values, achieved for different regularized regressions (RR, LSRR=CR(1), CR(2), CR(3), the best PLS, and the best PCR). To some degree of approximation, this feature is frequently seen. Note also that LSRR is less sensitive

than RR to the choice of the ridge constant.

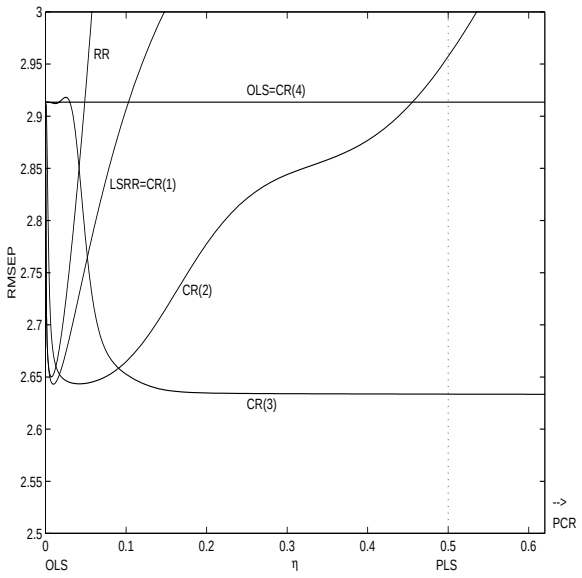

Figure 1: RMSEP for five different predictors: RR, LSRR and CR(k) for varying k. Data taken from [7], n=13, dim(x)=4. RMSEP is plotted against $\eta = \delta/(2\delta + \lambda_{max})$, $0 \leq \eta \leq 1$, for $\delta \geq 0$ or $< -\lambda_{max}$, where λ_{max} is the largest principal value.

An alternative continuum regression

A method suggested in [8] and mentioned in the discussion of [9], has also been called continuum regression, but has later been referred to as CPR (*continuum power regression*), e.g. in [6]. In this method the matrix $\mathbf{X}$ is transformed to a continuum power $\mathbf{X}^{(\gamma)}$, by raising the singular values of $\mathbf{X}$ to the power γ, followed by ordinary PLS with the new $\mathbf{X} = \mathbf{X}^{(\gamma)}$. Like CR, this method also encompasses OLS, PLS and PCR. However, the construction criterion does not depend solely on the variance and the correlation, hence CPR is not equivalent to CR.

Joint continuum regression

A natural question is that in the case of multivariate response variable y, can one gain efficiency by using some multivariate version of the univariate technique, instead of applying the latter separately on each component? Brooks & Stone [3] proposed a multivariate version of their CR, called *joint continuum regression* (JCR). Although they did not find it very promising in practice, it ties together several methods as special cases of JCR. The role of the OLS from the univariate situation is taken by the reduced rank regression (RRR), whereas PCR remains as the other limiting case. Multivariate PLSR appears in a form introduced by de Jong [4] under the name SIMPLS.

It is probably true that both the univariate and the multivariate versions of the continuum regression have not been so far as important for statistical practice per se than as a framework for tying up methods and for promoting a greater understanding of various methods and their intimate relationships. One observation is that the choice of regularized regression method seems to be of lesser importance than the choice of data pretreatment procedure, since all the methods suffer from the lack of scale invariance.

References

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

[2] Björkström, A. and Sundberg, R. (1999). A generalized view on continuum regression. *Scand. J. Statist.*, **26**, 17–30.

[3] Brooks, R. and Stone, M. (1994). Joint continuum regression for multiple predictands. *J. Amer. Statist. Assoc.*, **89**, 1374–1377.

[4] de Jong, S. (1993) SIMPLS: an alternative approach to partial least squares regression. *Chemometrics and Intell. Lab. Systems*, **18**, 251–263.

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

[6] de Jong, S., Wise, B. M. and Ricker, L. (2001). Canonical partial least squares and continuum power regression. *J. Chemometrics*, **15**, 85–100.

[7] Hald, A. (1952). *Statistical Theory with Engineering Applications.* Wiley, New York. (Data on page 647)

[8] Lorber, A., Wangen, L. E. and Kowalski, B. R. (1987). A theoretical foundation for the PLS algorithm. *J. Chemometrics*, **1**, 19–31.

[9] 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. Roy. Statist. Soc. Ser. B*, **52**, 237–269; corrigendum (1992) **54**, 906–907.

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

Further Reading

Brown, P. J. (1993). *Measurement, Regression, and Calibration.* Oxford University Press, Oxford. (Chapter 4 treats regularized regression)

Sundberg, R. (1999). Multivariate calibration—direct and indirect regression methodology (with discussion). *Scand. J. Statist.,* **26**, 161–207. (A review type paper)

Related Entries: Conditioning diagnostics, linear regression, shrinkage estimators, weak data