Shrinkage regression

Rolf Sundberg

Volume 4, pp 1994–1998

in

Encyclopedia of Environmetrics
(ISBN 0471 899976)

Edited by

Abdel H. El-Shaarawi and Walter W. Piegorsch

Shrinkage regression

Shrinkage regression refers to **shrinkage** methods of estimation or prediction in regression situations, useful when there is **multicollinearity** among the regressors. With a term borrowed from approximation theory, these methods are also called **regularization methods**. Such situations occur frequently in environmetric studies, when many chemical, biological or other explanatory variables are measured, as when Branco et al. [3] wanted to model toxicity of $n = 19$ industrial wastewater samples by pH and seven metal concentrations, and found it necessary to use dimensionality-reducing techniques. As another example, to be discussed further below, Andersson et al. [1] constructed a **quantitative structure–activity relationship** (QSAR) to model polychlorinated biphenyl (PCB) effects on cell-to-cell communication inhibition in the liver, having a sample of size $n = 27$ of PCB congeners that represented a set of 154 different PCBs, and using 52 physicochemical properties of the PCBs as explanatory variables.

Suppose we have n independent observations of $(\mathbf{x}, y) = (x_1, \ldots, x_p, y)$ from a standard multiple regression model

$$Y = \alpha + \beta'\mathbf{x} + \varepsilon \tag{1}$$

$\text{var}(\varepsilon) = \sigma^2$ (*see* **Linear models**). The ordinary **least squares** (OLS) estimator can be written

$$\mathbf{b}_{\text{OLS}} = \mathbf{S}_{\mathbf{xx}}^{-1}\mathbf{s}_{\mathbf{x}y} \tag{2}$$

where $\mathbf{S}_{\mathbf{xx}}$ is the sum of squares and products matrix of the centered $\mathbf{x}$ variables and $\mathbf{s}_{\mathbf{x}y}$ is the vector of their sums of products with y. The OLS estimator is the best fitting and minimum variance linear unbiased estimator (best linear unbiased estimator, BLUE), with variance

$$\text{var}(\mathbf{b}_{\text{OLS}}) = \sigma^2\mathbf{S}_{\mathbf{xx}}^{-1} \tag{3}$$

That OLS yields the best fit does not say, however, that $\mathbf{b}_{\text{OLS}}$ is best in a wider sense, or even a good choice. We will discuss here alternatives to be used when the $\mathbf{x}$ variables are (near) multicollinear, that is when there are linear combinations among them that show little variation. The matrix $\mathbf{S}_{\mathbf{xx}}$ is then near singular, so $\text{var}(\mathbf{b}_{\text{OLS}})$ will have very large elements. Correspondingly, the components of $\mathbf{b}_{\text{OLS}}$ may show unrealistically large values. Under exact collinearity, $\mathbf{b}_{\text{OLS}}$ is not even uniquely defined. In these situations it can pay substantially to use shrinkage methods that trade bias for variance. Not only can they give more realistic estimates of β, but they are motivated even stronger for construction of a predictor of y from $\mathbf{x}$.

A simple and as we will see fundamental shrinkage construction is *ridge regression* (RR), according to which the OLS estimator is replaced by

$$\mathbf{b}_{\text{RR}} = (\mathbf{S}_{\mathbf{xx}} + \delta\mathbf{I})^{-1}\mathbf{s}_{\mathbf{x}y} \tag{4}$$

for a suitable *ridge constant* $\delta > 0$ (to be specified). Addition of whatever $\delta > 0$ along the diagonal of $\mathbf{S}_{\mathbf{xx}}$ clearly makes this factor less close to singular. To see more precisely how RR shrinks, consider an orthogonal transformation that replaces the components of $\mathbf{x}$ by the principal components (PCs), $t_i = \mathbf{c}_i{}'\mathbf{x}$, $i = 1, \ldots, p$. Here the coefficient vectors $\mathbf{c}_i$ are the eigenvectors of $\mathbf{S}_{\mathbf{xx}}$, with eigenvalues λ_i, say. Correspondingly, $\mathbf{S}_{\mathbf{xx}}$ is replaced by the diagonal matrix of these eigenvalues λ_i, and RR adds δ to them all. When OLS divides the ith component of the vector $\mathbf{s}_{ty}$ by λ_i, RR divides by $\lambda_i + \delta$. Hence, RR shrinks $\mathbf{b}_{\text{OLS}}$ much in principal directions with small λ, but only little in other principal directions.

Variable selection might appear to shrink, by restricting attention to a lower-dimensional subspace U_r $(r < p)$ of the model space of the $\mathbf{x}$s. Variable selection improves precision when it omits variables with negligible influence, but if it is used to eliminate multicollinearities, then it does not necessarily shrink and will make interpretation difficult. More sophisticated ways of restricting attention to lower-dimensional subspaces U_r of regressors are represented by methods such as *principal components regression* (PCR) and *partial least squares* (or *projection to latent structures*; PLS) regression. Both methods construct new regressors, linear in the old regressors, and y is regressed (by least squares, LS) on a relatively small number of these PCs or PLS factors. This number r, as well as the ridge constant δ in RR, is typically chosen by a **cross-validation** criterion. As an example, Andersson et al. [1] used PLS regression with two PLS factors to replace the 52 original variables in the regression. This model explained $R^2 = 84\%$ of the variability, and in a cross-validation predicted 80% (necessarily less than R^2, but quite close in this case). They could also interpret their model within the context of their chemometric application, as referring basically to the position pattern of the chlorine atoms.

In the standard form of PCR, the new regressors $t_i = \mathbf{c}_i{}'\mathbf{x}$ are the first PCs of the (centered) $\mathbf{x}$ from a PCs analysis. This means that the $\mathbf{c}_i$s are the eigenvectors of $\mathbf{S_{xx}}$ that explain as much variation as possible in $\mathbf{x}$. The space U_r is then spanned by the first r PCs. In this way directions in $\mathbf{x}$ with little variation are shrunk to 0. The connection with variation in y is only indirect, however. PLS regression is more efficient in this respect, because it constructs a first regressor $t_1 = \mathbf{c}_1{}'\mathbf{x}$ to maximize *covariance* between t_1 and y (among coefficient vectors $\mathbf{c}_1$ of fixed length), and then successively t_2, t_3, etc. so as to maximize covariance with the y residuals after LS regression of y on the previous set of regressors. As an algorithm for construction of the PLS regression subspaces U_r of the $\mathbf{x}$ space, spanned by $t_1, \ldots t_r$, this is one of several possible. For a long time PLS was only algorithmically defined, until Helland [10] gave a more explicit characterization by showing that U_r for PLS is spanned by the sequence of vectors $\mathbf{s_x}_y$, $\mathbf{S_{xx}s_x}_y, \ldots, \mathbf{S_{xx}^{r-1}s_x}_y$. This representation is of more theoretical than practical interest, however, because there are numerically better algorithms.

A step towards a general view on the choice of shrinkage regression was taken by Stone and Brooks [13], who introduced *continuum regression* (CR), with the criterion function

$$R^2(t, y)V(t)^\gamma \tag{5}$$

where $R(t, y)$ is the sample correlation between $t = \mathbf{c}'\mathbf{x}$ and y, $V(t)$ is the sample variance of t, and $\gamma \ge 0$ is a parameter to be specified. This function of t (or $\mathbf{c}$) should be maximized under the constraint $|\mathbf{c}| = 1$. In previous terms

$$R^2(t, y) \propto \frac{(\mathbf{c}'\mathbf{s_x}_y)^2}{\mathbf{c}'\mathbf{S_{xx}c}} \tag{6}$$

$$V(t) \propto \mathbf{c}'\mathbf{S_{xx}c} \tag{7}$$

As in PLS, when a first $\mathbf{c} = \mathbf{c}_1$, $t = t_1$, has been found, the criterion is applied again on the residuals from LS regression of y on t_1 to find the next $\mathbf{c}$, $\mathbf{c} = \mathbf{c}_2$, etc. The new $\mathbf{c}$ will automatically be orthogonal to the previous ones. In order to institute a regression method of this algorithm, the number r of regressors and the parameter γ are chosen so as to optimize some cross-validation measure.

The most important role of CR is perhaps not as a practical regression method, however, but as a theoretical device embracing several methods. For $\gamma = 0$ we get back OLS, and the algorithm will stop after one step ($r = 1$). For $\gamma = 1$ the criterion function (5) is proportional to the covariance squared and will yield PLS, and as $\gamma \to \infty$, $V(t)$ will dominate and we get PCR. A much more general yet simple result, which relates to RR and provides an even wider understanding, has been given in [2], as follows.

Suppose we are given a criterion that is a function of $R^2(t, y)$ and $V(t)$, nondecreasing in each of these arguments. The latter demand is quite natural, because we want both of these quantities to have high values. The regressor maximizing any such criterion under the constraint $|\mathbf{c}| = 1$ is then proportional to the RR estimator (4) for some δ, possibly negative or infinite. Conversely, for each $\delta \ge 0$ or $\delta < -\lambda_{\max}$ in (4) there is a criterion for which this is the solution. The first factor PLS is obtained as $\delta \to \pm\infty$, and $\delta \to -\lambda_{\max}$ corresponds to the first factor PCR. The first factor CR with varying $\gamma \ge 0$ usually (but not always) requires the whole union of $\delta \ge 0$ and $\delta < -\lambda_{\max}$. The use of $\mathbf{c} = \mathbf{b}_{\mathrm{RR}}$ from (4) to form a single regressor in LS regression is called *least squares ridge regression* (LSRR) in [2], and differs from RR by a scalar factor that will be near 1 when δ is small. Hence RR is closely related to the other constructions. Furthermore, experience shows that typically RR and LSRR for some best small δ yield a similar regression as (the best r) PLS and (best r) PCR.

If different constructions yield quite similar estimators/predictors, what shrinkage method is then preferable? Here is some general advice.

- RR and LSRR have the advantage that they only involve a single factor.
- LSRR only shrinks to compensate for multicollinearity, whereas RR always shrinks.
- RR has a Bayesian interpretation as posterior mean for a suitable prior for β (*see* **Bayesian methods and modeling**).
- LSRR (like PLS and PCR) yields residuals orthogonal to the fitted relation, unlike RR.
- PLS and PCR have interpretations in terms of latent factors, such that for PCR essentially all $\mathbf{x}$ variability is in U_r for suitable r, whereas for PLS all $\mathbf{x}$ variability that influences y is in the corresponding U_r for a typically somewhat smaller r than for PCR.

- The latent factor structure in PCR and PLS is convenient for **outlier** detection and **classification** (cf. below and [12, Chapter 5]). Also **x** components missing at random in prediction of y for a new observation are easily handled.

Multicollinearity imposes estimation identifiability problems, when we want to find the true β. If $x_1 + x_2$ is (near) constant, say, we cannot tell if x_1 or x_2 or both jointly are responsible for the influence of **x** on y. This is less serious for prediction, provided that new observations satisfy the same relation between y and **x**, and that new **x**s are like calibration **x**s, in particular satisfying the same multicollinearity. To judge the predictive ability, some version of cross-validation is natural, based on repeated splits of data in calibration and test sets. But we must always have in mind the risk that the calibration (**x**, y) data are particular in some respect, and that therefore cross-validation criteria are too optimistic. An extreme example, from a study concerning determination of nitrogen in wastewater, is discussed in [14]. In the QSAR example above [1], the model was fitted on a sample of 27 PCBs from a population group of 154 PCBs. If we want to predict the activity y for a new PCB from this population, then we must allow and check for the possibility that it is an outlier in **x** space, compared with the calibration sample. Implications differ according to whether it is an outlier in a direction orthogonal to the two-dimensional latent subspace U_r, or only within U_r. There are also other PCBs not even satisfying the criteria for belonging to the population group under study, and for them we must doubt the prediction equation for y even if **x** is not an outlier.

Shrinkage regression constructions are intrinsically scale noninvariant. This is because ‘near’ in near-collinear is metric-dependent. Correlation is invariant, but not covariance and variance. Individual rescaling of the **x** components, or any other nonorthogonal transformation, will affect the shrinkage. It is often recommended that **x** components be autoscaled to unit variances as a pretreatment. This is natural when the **x** components are on incomparable scales, as in QSAR studies, but silly in many other situations, for example in spectroscopy. As another example of frequent pretreatments, second-order differencing of a spectrum **x** corresponds to a nonorthogonal linear transformation, and thus has effects on shrinkage regression. The choice of scaling is often more crucial than the choice between the different methods (RR, LSRR, PCR, PLS).

In multivariate regression, when the response is a vector **y** and not a scalar y, and the regression coefficients form a matrix **B**, we could try shrinkage also in **y** space. In this way we might be able to borrow strength for the prediction of one component of **y** from the others. A variety of methods has been proposed in the literature.

There are PLS type algorithms (PLS2, SIMPLS), in each step constructing a pair of PLS factors, in **y** and in **x**. The first factors are selected equally by PLS2 and SIMPLS, but their criteria for the subsequent factors differ. There is a multivariate version of CR, *joint continuum regression* (JCR [5]), of which SIMPLS is a special case. At one of the extreme ends of JCR we find PCR, which is not affected by the dimension of **y**. At the other end is the LS version of **reduced-rank regression** (RRR). RRR, imposing a rank constraint on **B**, is an example of a principle that essentially only shrinks in **y** space and therefore is inadequate under near-collinearity in **x**. The maximum likelihood (ML) ($\neq$ LS) estimator in an RRR model (standard multivariate regression with a fixed rank constraint added) is equivalent to using a number of canonical variates from a canonical correlation analysis, so for each canonical y variate we use OLS regression. This method has become an important tool in econometrics for modeling cointegration (i.e. common trends) in nonstationary vector autoregressive **time series** models [11], with potential applications also in **environmetrics**.

Burnham et al. [8] formulated a latent variable multivariate regression model, in which both y and x have latent structures, more or less overlapping. The overlap yields a regression relationship, when **y** should be predicted from **x**. A constrained ML method in this model is equivalent to a method known as *principal covariates regression* (PcovR [9]), which is a shrinkage method that will often yield similar results to those of PLS2 and SIMPLS. See also [7], giving characterizations and relationships between several different methods with respect to their optimality criteria of construction.

Further methods were discussed and compared by Breiman and Friedman [4], who constructed and advocated a method they called ‘*curds and whey*’, primarily shrinking in **y** space. The article caused lively debate and much criticism.

Despite the many shrinkage methods proposed and used for multivariate regression, it is not clear if and when they are really better than separate univariate regressions. From the literature it seems as if one seldom gains much precision from replacing separate univariate predictions by a multivariate prediction method, cf. the discussion in [5].

For fully implemented PCR or PLS, with cross-validation to help select the number of factors and various diagnostic tools, there are commercial packages from the field of **chemometrics** doing the job, such as SIMCA (Umetrics) and UNSCRAMBLER® (CAMO). For those who want more flexibility or more lucid software, programming in Matlab® can be recommended. The PLS Toolbox (Eigenvector Research) contains a large number of Matlab subroutines, and there is also much free code available on the web.

Recommended literature on shrinkage regression includes the books by Brown [6], not least for covering Bayesian aspects, and Martens and Næs [12].

References

[1] Andersson, P., Haglund, P., Rappe, C. & Tysklind, M. (1996). Ultra violet absorption characteristics and calculated semi-empirical parameters as chemical descriptors in multivariate modelling of polychlorinated biphenyls, *Journal of Chemometrics* **10**, 171–185.

[2] Björkström, A. & Sundberg, R. (1999). A generalized view on continuum regression, *Scandinavian Journal of Statistics* **26**, 17–30.

[3] Branco, J.A., Pires, A.M., Mendonça, E. & Picado, A.M. (1994). A statistical evaluation of toxicity tests of waste water, in *Statistics for the Environment*, Vol. 2, V. Barnett & K.F. Turkman, eds, Wiley, Chichester.

[4] Breiman, L. & Friedman, J.H. (1997). Predicting multivariate responses in multiple linear regression (with discussion), *Journal of the Royal Statistical Society, Series B* **59**, 3–54.

[5] Brooks, R.J. & Stone, M. (1994). Joint continuum regression for multiple predictands, *Journal of the American Statistical Association* **89**, 1374–1377.

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

[7] Burnham, A.J., Viveros, R. & MacGregor, J.F. (1996). Frameworks for latent variable multivariate regression, *Journal of Chemometrics* **10**, 31–45.

[8] Burnham, A.J., MacGregor, J.F. & Viveros, R. (1999). A statistical framework for multivariate latent variable regression methods based on maximum likelihood, *Journal of Chemometrics* **13**, 49–65.

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

[10] Helland, I.S. (1988). On the structure of partial least squares regression, *Communications in Statistics–Simulation and Computation* **17**, 581–607.

[11] Johansen, S. (1995). *Likelihood-based Inference in Cointegrated Vector Autoregressive Models*, Oxford University Press, Oxford.

[12] Martens, H. & Næs, T. (1989). *Multivariate Calibration*, Wiley, Chichester.

[13] 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), *Journal of the Royal Statistical Society, Series B* **52**, 237–269.

[14] Sundberg, R. (1999). Multivariate calibration – direct and indirect regression methodology (with discussion), *Scandinavian Journal of Statistics* **26**, 161–207.

(*See also* **Collinearity**)

ROLF SUNDBERG