Cross-validated Regression Estimation

Roy E. Welsch
Massachusetts Institute of Technology, Sloan School of Management
One Amherst Street, E40-129
Cambridge, MA 02142, USA
rwelsch@mit.edu

In Welsch (2000) we compared a variety of ridge type regression estimators with partial least squares (PLS). Our results confirmed those of Frank and Friedman (1993) who showed that PRESS cross-validated ridge regression (CVRR) performed better than cross-validated PLS and any of the estimators considered in either paper. One estimator (a modification of the Hoerl and Kennard generalized ridge estimator) performed very well without cross-validation in Welsch (2000) and was second only to CVRR. In this paper we consider additional estimators of this type.

To simplify notation, we will use the singular value decomposition (SVD) of $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathrm{T}}$ with $\mathbf{U}$ $n \times p$, $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}$, $\mathbf{V}$ a $p \times p$ orthogonal matrix, and $\mathbf{D}$ a $p \times p$ diagonal matrix with non-increasing diagonal entries. Since the estimators we are considering are equivariant, we will change to the rotated coordinate system $\mathbf{Z} = \mathbf{X}\mathbf{V}$ and the generalized ridge estimator is

$$(1) \qquad \hat{\boldsymbol{\alpha}}_{\mathrm{GR}} = \left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z} + \mathbf{K}\right)^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{y} = \left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z} + \mathbf{K}\right)^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{Z}\hat{\boldsymbol{\alpha}}_{\mathrm{LS}}$$

where $\mathbf{K}$ is a $p \times p$ diagonal matrix with diagonal elements k_j and $\hat{\boldsymbol{\alpha}}_{\mathrm{LS}}$ is $\hat{\boldsymbol{\alpha}}_{\mathrm{GR}}$ with all $k_j = 0$.

Using the SVD, this becomes

$$(2) \qquad \hat{\boldsymbol{\alpha}}_{\mathrm{GR}} = \mathrm{diag}\left[r_j\right]\hat{\boldsymbol{\alpha}}_{\mathrm{LS}}, \qquad r_j = d_j^2 / (d_j^2 + k_j)$$

and the r_j are the "shrinkage" factors. Note that in the $\mathbf{Z}$ coordinate system the t-statistics for the least-squares coefficients $\hat{\boldsymbol{\alpha}}_{\mathrm{LS}}$ are just $t_j = (\mathbf{u}_j^{\mathrm{T}}\mathbf{y})/s$ with $\mathbf{u}_j$ denoting the jth column of $\mathbf{U}$ and s the standard error of the least-squares regression.

The Hoerl and Kennard generalized ridge estimator (see Gruber, 1988) uses $k_j = s^2 \Big/ \left(\hat{\alpha}_{\mathrm{LS}}\right)_j^2$

which gives

$$(3) \quad \text{(HK)} \qquad r_j = t_j^2 / (t_j^2 + 1).$$

In Welsch (2000), we found that using

$$(4) \quad \text{(HKA)} \qquad r_j = \left(\frac{t_j^2}{t_j^2 + 1}\right)^{c_j}$$

where $c_j = \lambda_{\max} / \lambda_j$ and $\lambda_j = d_j^2$ are the eigenvalues of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ was a good choice and was beaten only by CVRR. Since HKA is not a cross-validated procedure, it is easier to compute than CVRR (and PLS). This estimator was motivated by the discussion in Thorpe and Scharf (1995). They noted that many ridge estimators tend to have a value of r_j that is too large for directions with small λ_j and

that values near 0.9 for $t_j^2 / (t_j^2 + 1)$ might be required before the relationship of the response variable, **y**, to the principal components, $\mathbf{z}_j$, should begin to override the general rule of essentially removing principal components with small λ_j. However, it is natural to wonder what would happen if we did cross-validate HKA. We tested the following functional forms:

(5) $r_j = f\left(t_j^2\right)^{c_j}$ with $f_1(x) = (x + \gamma) / (x + \gamma + 1)$, $f_2(x) = \gamma x / (\gamma x + 1)$, $f_3(x) = x^\gamma / (x^\gamma + 1)$,

or

$$r_j = \left(t_j^2 \Big/ \left(t_j^2 + 1\right)\right)^{g\left(c_j\right)} \text{ with } g_1(c_j) = c_j + \gamma, \quad g_2(c_j) = \gamma c_j, \quad g_3(c_j) = c_j^\gamma.$$

The parameter γ will be chosen by cross-validation.

We used the same Monte Carlo procedures as described in Welsch (2000) which set $n = 50$, $p = 5$ or 40, with a variety of choices for the signal-to-noise ratio, correlation structure of $\mathbf{X}^T\mathbf{X}$, and settings for the true α_j. For each estimator, the average squared prediction error (over 100 simulations) was:

CVRR	.6359	LS	1.4437
HKA	.7463	PLS	.8110
F1	.8165	G1	.7565
F2	.8062	G2	.7250
F3	.7459	G3	1.0998

To our surprise, only G2 made any real improvement. This attempts to let cross-validation choose whether $\lambda_{\max}$ was the right numerator for $\lambda_{\max} / \lambda_j$. None beat CVRR, but many bested PLS. It appears that bringing information about t_j^2 into the generalized ridge parameters works fairly well without cross-validation, but does not pay off when cross-validation is available.

REFERENCES

Frank, I.E. and Friedman, J.H. (1993), "A Statistical View of Some Chemometrics Regression Tools," *Technometrics*, 35, 109-148.

Gruber, M.H.J. (1998), *Improving Efficiency by Shrinkage*, New York: Marcel Dekker.

Thorpe, A.J. and Scharf, L.L. (1995), "Data Adaptive Rank-Shaping Methods for Solving Least Square Problems," *IEEE Transactions on Signal Processing*, 43, 1591-1601.

Welsch, R.E., "Is Cross-Validation the Best Approach for Principal Component and Ridge Regression?" To appear in the *Proceedings of the 32nd Symposium on the Interface: Computing Science and Statistics.*

RESUME

Nous considérons de nouveaux estimateurs de type régression ridge, en nous inspirant de l'analyse Monte-Carlo de Welsch (2000) et des développement théoriques de Thorpe and Scharf (1995).