

Is Cross-Validation the Best Approach for Principal Component and Ridge Regression?

Roy E. Welsch
Massachusetts Institute of Technology
Cambridge, MA 02139
rwelsch@mit.edu

Abstract

A recent study by Frank and Friedman (1993) indicated that cross-validated ridge regression performed well when compared to partial least-squares regression and cross-validated principal components regression. Thorpe and Scharf (1995) consider a number of uncross-validated ridge-type estimators from an engineering point of view. In this paper we examine a variety of estimators to see if we can do as well as or nearly as well as fully cross-validated ridge regression. We conclude that when the number of parameters does not exceed the number of observations, it may be possible to avoid cross-validation.

1 Introduction

For many years, we have pointed to the Frank and Friedman (1993) paper as evidence that cross-validated principal component regression (CVPCR) and cross-validated ridge regression (CVRR) were worthy competitors to cross-validated partial least-squares or partial latent structure regression (PLS). CVPCR and CVRR are computationally simpler and, generally, easier to understand than PLS. PLS retains a large following, perhaps because the cross-validated form is widely available, while only non-cross-validated forms of PCR and RR are widely available, and these are generally not as good as PLS. An overview of PLS from a statistician's point of view may be found in Garthwaite (1994).

The more recent appearance in the engineering literature of a paper by Thorpe and Scharf (1995), while not a direct comparison with PLS, provided a wide range of non cross-validated ridge-like regression estimators that could be compared with those studied in the Frank and Friedman (1993) paper.

Of course, there is a vast literature on ridge regression and we will make no attempt to address all of the proposed estimators here. However, Krzanowski and Marriott (1995) suggest that PC regression using t -statistics to decide which PC's should be in the model might be a good PLS competitor because information about the response variable

is being used to decide which components are in the regression. This has long been noted as a plus for PLS and a minus for standard PC regression where no response variable information is used to choose the principal components retained. It is important to note that CVPCR does use response variable information during the cross-validation process. Rawlings in his textbooks (Rawlings 1989, 1998) has suggested that standard PCR regression should be modified by using the t -statistics so that principal components with large eigenvalues or large t -statistics are included in the regression.

Our goal in this paper is to compare some of the regression estimators suggested above along with a few others that surfaced during the simulation study. The outline of the paper is as follows. In Section 2, we develop the general notion of rank-adjusted estimators. In Section 3, we discuss the various estimators in a uniform notation. In Section 4, we describe our simulation models which follow closely those used in the Frank and Friedman (1993) paper. In the last section, we provide our conclusions and recommendations.

2 Background and Notation

We assume that we have n training observations with $p + 1$ parameters (the intercept will always be included) and a single response variable. We shall also assume that all of the data are standardized by subtracting means and dividing by the standard deviation (with n , not $(n - 1)$ as the divisor). Many of the estimators are not scale invariant, so this scaling will affect the analysis. Thus our standardized model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

with \mathbf{y} an $n \times 1$ vector, \mathbf{X} and $n \times p$ matrix, $\boldsymbol{\beta}$ a $p \times 1$ vector and $\boldsymbol{\epsilon}$ an $n \times 1$ vector. To simplify our notation (and computation, in many cases), we will use the singular value decomposition of \mathbf{X} , $\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T$ with \mathbf{U} an $n \times n$ orthogonal matrix, \mathbf{V} a $p \times p$ orthogonal matrix, and \mathbf{S} an $n \times p$ diagonal matrix with non-increasing diagonal entries. To address cases where $p > n$, we use the pseudo-inverse of \mathbf{X} , denoted by \mathbf{X}^+ ,

$$\mathbf{X}^+ = \mathbf{V}\mathbf{S}^+\mathbf{U}^T.$$

All of the estimators we will consider, except PLS, will modify \mathbf{X} in the following way:

$$\begin{aligned}\mathbf{X}_r &= \mathbf{U}\mathbf{S}_r\mathbf{V}^T \\ \mathbf{S}_r &= \mathbf{R}^+\mathbf{S}\end{aligned}$$

where \mathbf{R} is an $n \times n$ diagonal matrix with at most $\min(n, p)$ nonzero diagonal elements.

In traditional least-squares estimation,

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\text{LS}} &= \mathbf{X}^+ \mathbf{y} = \mathbf{V}\mathbf{S}^+\mathbf{U}^T\mathbf{y} \\ &= \sum_j^{\min(p,n)} \mathbf{v}_j \frac{1}{s_j} \mathbf{u}_j^T \mathbf{y}\end{aligned}$$

where u_j and v_j are the appropriate columns of \mathbf{U} and \mathbf{V} . When $p > n$, this is the minimum norm solution to the least-squares equations. Rank-adjusted estimation gives

$$\hat{\boldsymbol{\beta}}_r = \mathbf{X}_r^+ \mathbf{y} = \sum_j \mathbf{v}_j \frac{r_j}{s_j} \mathbf{u}_j^T \mathbf{y}$$

where the r_j are the diagonal elements of \mathbf{R} . Since the estimators we will consider are equivariant, we will change to the rotated coordinate system, $\mathbf{Z} = \mathbf{X}\mathbf{V}$ with uncorrelated explanatory variables with coefficients, $\boldsymbol{\alpha} = \mathbf{V}^T\boldsymbol{\beta}$.

Traditional ridge regression is

$$\begin{aligned}\hat{\boldsymbol{\alpha}}_{\text{RR}} &= (\mathbf{Z}^T\mathbf{Z} + k\mathbf{I})^{-1} \mathbf{Z}^T \mathbf{y} \\ &= \sum_j \frac{s_j}{s_j^2 + k} \mathbf{u}_j^T \mathbf{y} \\ &= \sum_j \frac{r_j}{s_j} \mathbf{u}_j^T \mathbf{y}\end{aligned}$$

with $r_j = s_j^2 / (s_j^2 + k)$. We also note that

$$\begin{aligned}\hat{\boldsymbol{\alpha}}_{\text{RR}} &= (\mathbf{Z}^T\mathbf{Z} + k\mathbf{I})^{-1} \mathbf{Z}^T \mathbf{Z} \hat{\boldsymbol{\alpha}}_{\text{LS}} \\ &= \text{diag} \left[\frac{s_j^2}{s_j^2 + k} \right] \hat{\boldsymbol{\alpha}}_{\text{LS}}.\end{aligned}$$

Thus r_j is the ‘‘shrinkage’’ factor, $s_j^2 / (s_j^2 + k)$, associated with many forms of biased regression. Finally, we note that the t -statistic for $(\hat{\boldsymbol{\alpha}}_{\text{LS}})_j$ is just $t_j = \mathbf{u}_j^T \mathbf{y} / \sigma$. When σ is known $E(t_j) = (s_j \alpha_j) / \sigma$, but since $t_j \sim N((s_j \alpha_j) / \sigma, 1)$, $E(t_j^2) =$

$s_j^2 \alpha_j^2 / \sigma^2 + 1$ and is, therefore, biased. Hence we will use $t_j^2 - 1$ in some situations.

3 The Estimators

We have chosen to rely on the Frank and Friedman (1993) study for comparisons of PLS with other estimators such as CVPCR and CVRR. These later two along with LS have been carried along in our study (which uses a simulation structure like that in Frank and Friedman (1993)) in order to allow comparison with PLS.

Our study includes a number of additional biased estimators that were not considered in the Frank and Friedman (1993) paper. Many of these come from the Thorpe and Scharf (1995) paper, but were not necessarily first suggested by those authors.

3.1 Least-squares

This is the traditional approach. When $p > n$, and the solution is not unique, the minimum norm solution is used. In this situation, all residuals are zero and estimation of σ^2 is impossible.

3.2 Ridge Regression

Most of the estimators we will consider are of this form. To simplify notation, let $\lambda_j = s_j^2$. These are two basic types, those with one ridge parameter

$$r_j = \lambda_j / (\lambda_j + k),$$

and those with more than one (generalized ridge)

$$r_j = \lambda_j / (\lambda_j + k_j).$$

In problems with p near to or greater than n , we can suspect some difficulties because we are trying to estimate a large number of model parameters, like $\boldsymbol{\alpha}$, and of pseudo-parameters, like the k_j . In some cases, the $p > n$ problem cannot be considered at all because of difficulties estimating σ .

Cross-validated ridge regression (CVRR) finds the single k by cross-validation using the PRESS statistic as described in Frank and Friedman (1993). It proved to be the ‘‘best’’ estimator considered in that study. It is easily applied in both the $n \geq p$ and $n < p$ cases. To see how much would be lost using generalized cross validation (GCV) as described in Golub, et. al. (1979) instead of PRESS, we also chose the k in ridge regression using GCV and called the result GCVR.

Can we do better? Can we do well without cross-validation? Under the usual model conditions, we can compute the theoretical mean square error (MSE) and mean square prediction error (MSPE) for rank-adjusted estimators. Then these quantities can be minimized and the appropriate k or k_j found. Many of the details are in Thorpe and Scharf (1995)

Using these ideas, Thorpe and Scharf suggest the following biased and unbiased estimators for r_j when there is more than one k_j :

$$(HK) \quad r_j = \frac{(\mathbf{u}_j^T \mathbf{y})^2}{\left[(\mathbf{u}_j^T \mathbf{y})^2 + \sigma^2 \right]} \\ = t_j^2 / (t_j^2 + 1)$$

$$(HKUB) \quad r_j = \max \left(0, \frac{(\mathbf{u}_j^T \mathbf{y})^2 - \sigma^2}{(\mathbf{u}_j^T \mathbf{y})^2} \right) \\ = \max \left(0, (t_j^2 - 1) / t_j^2 \right)$$

with $t_j^2 = (\mathbf{u}_j^T \mathbf{y})^2 / \sigma^2$. The estimator HK reduces to $k_j = \sigma^2 / (\hat{\alpha}_{LS})_j^2$ which is the original Hoerl and Kennard (1970) estimate. HKUB has also been suggested in various forms, see Gruber (1998) for references.

Thorpe and Scharf also point out that HK and similar estimators probably give too much weight to the t -statistics and fail to shrink enough when λ_j is small. We note that for HK,

$$r_j = \frac{t_j^2}{t_j^2 + 1} = \frac{1}{1 + \frac{1}{t_j^2}}$$

and when t_j^2 is large, the component is essentially fully retained independent of the size of λ_j . Therefore we tried to adjust for this by considering

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

where $c_j = \lambda_{\max} / \lambda_j$. This means that when c_j is large (λ_j relatively small), there is considerably more shrinkage than would be the case with just t_j^2 . The "unbiased" form is just

$$(HKUBA) \quad r_j = \left[\frac{\max(0, t_j^2 - 1)}{t_j^2} \right]^{c_j}.$$

Throughout their paper, Thorpe and Scharf assume that σ^2 is known. When $p < n$, we will estimate σ^2 by using residuals based on the LS solution. For $p \geq n$, this is not feasible and σ^2 becomes a nuisance parameter that could possibly be obtained by cross-validation.

Finally, Thorpe and Scharf suggest a conditional mean estimator of the r_j which we will call TSCM:

$$(TSCM) \quad [r_j]_{CM} = E[r_j | \mathbf{y}]$$

This estimator requires numerical quadrature for each r_j and needs an estimate of σ^2 / σ_s^2 where σ_s^2 is the common prior variance of the α parameters. As is done in ridge regression, this ratio can be found by cross-validation. An estimate of σ^2 is also required. For the problems we considered $n = 50$, $p = 5, 40$, cross-validation and even generalized cross-validation took far too long to be practical. We were only able to run a few simulations and the results were not encouraging even though, theoretically, TSCM has many interesting properties.

Estimates like those suggested by Thorpe and Scharf can be obtained under the constraint that

$$r_j = \frac{\lambda_j}{\lambda_j + k}$$

where k is constant across j . Four estimators arise based on MSE, MSPE and the unbiased forms MSEUB and MSPEUB. These quantities are minimized over k and lead to solving for k the first order equations:

$$(MSE) \quad \sum_j \frac{kt_j^2 - \lambda_j}{(\lambda_j + k)^3} = 0$$

$$(MSPE) \quad \sum_j \frac{\lambda_j (kt_j^2 - \lambda_j)}{(\lambda_j + k)^3} = 0$$

$$(UBMSE) \quad \sum_j \frac{kt_j^2 - \lambda_j - k}{(\lambda_j + k)^3} = 0$$

$$(UBMSPE) \quad \sum_j \frac{\lambda_j (kt_j^2 - \lambda_j - k)}{(\lambda_j + k)^3} = 0.$$

The UBMSPE is exactly the equation needed to find the optimal k for the C_k statistic suggested by Mallows (1974).

Of course, σ^2 is not known and for $p < n$, we use the LS estimate for σ^2 .

Some direct estimates of k have been suggested by analogy to the James-Stein estimator. Perhaps the most common is

$$(HKB) \quad k_{HKB} = \frac{p\sigma^2}{\hat{\mathbf{a}}_{LS}^T \hat{\mathbf{a}}_{LS}} = \frac{p}{\sum_{j=1}^p \lambda_j t_j^2}$$

due to Hoerl, Kennard and Baldwin (1975). Lawless and Wong (1976) suggested

$$(LW) \quad k_{LW} = \frac{p}{\sum_{j=1}^p t_j^2}$$

Since

$$\frac{1}{k_{HKB}} = \frac{1}{p} \sum_{j=1}^p \frac{1}{(k_{HKB})_j}$$

and

$$\frac{1}{k_{LW}} = \frac{1}{p} \sum_{j=1}^p t_j^2$$

we will also include $1/t_j^2$ as values for k_j in our study (GLW). An unbiased form would be $1/[\max(0, t_j^2 - 1)]$, which we will call GLWUB. A more complete discussion may be found in Gruber (1998), which, however, does not reference either the Frank and Friedman or the Thorpe and Scharf papers.

It is useful to note that when $k_j = 1/t_j^2$

$$r_j = \frac{\lambda_j}{\lambda_j + \frac{1}{t_j^2}}$$

allowing the size of the λ_j to play a role, at least for the moderate values of t_j^2 . This was not the case for HK.

3.3 Principal Components

Traditional PCR is done by ordering the components (orthogonal “explanatory” variables) according to the size of the eigenvalues. Components with “large” eigenvalues remain in the model. Those with “small” eigenvalues are removed from the model. Small may be made relative to large by looking at λ_{\max}/λ_j . Note that $\lambda_{\max}/\lambda_{\min}$ is the

condition index of $\mathbf{X}^T \mathbf{X}$. Cross validation avoids decisions about what is small or relatively small and appears to be the PC procedure of choice. Certainly, when comparing PLS to PCR as many people do, it seems only fair to compare CVPCR with PLS, since PLS is always cross-validated.

Krzanowski and Marriott (1995) and other authors have discussed the idea of ordering the principal components by size of t_j^2 rather than λ_j . This would be the in or out approach that is consistent with the ridge $t_j^2/(t_j^2 + 1)$ approach. Cross validation helps decide when the t_j^2 are large enough and adjusts for problems with multiple comparisons. We call this the CVTPCR.

Using t_j^2 -statistics in descending order is, essentially, forward stepwise regression in the PC space and the Frank and Friedman results indicate that regular cross-validated forward stepwise regression (VSS) is not necessarily a good procedure.

Rawlings (1988,1998) observes that a problem with traditional PC regression is that all decisions about which variables to retain are based on the explanatory variable data. He argues that a good working rule is to eliminate only those principal components that:

1. have small enough eigenvalues, and
2. for which the estimated LS regression coefficient $\hat{\alpha}_j$ is not significantly different from zero.

Rawlings suggests an eigenvalue is small when $\lambda_{\max}/\lambda_j > 100$ and that $\hat{\alpha}_j$ is not different from zero when the t -statistic fails to be significant at $\alpha = .10$ or $.20$. We call this procedure RPCR and use the $\alpha = .10$ value. A very conservative approach would be to only allow principal components with t significant at $\alpha = .10$ and $\lambda_{\max}/\lambda_j < 100$. We will call this WPCR. In order to mimic classical (uncross-validated) principal components regression we included CPCR which removes principal components with $\lambda_{\max}/\lambda_j > 100$. Again, we could try cross-validation to replace the λ_{\max}/λ_j and/or the α threshold. Double cross-validation will not be addressed in this paper.

4 Simulation Procedures

We attempted to follow the Monte Carlo experiments in Frank and Friedman as closely as possible to allow direct comparison. In particular, the number of training set observations was 50 and the number of explanatory variables was 5, 40, and 100 (where possible). The signal-to-noise ratio was $[\text{var}(\boldsymbol{\alpha}^T \mathbf{z})]^{1/2} / \sigma = 7, 3, 1$ and the population predictor-variable correlation matrix was either an identity diagonal matrix or the same matrix but with .9 placed in all of the off-diagonal elements. Finally, all $\alpha_j = 1$

or $\alpha_j = j^2$ to provide an unbalanced case. These factors provide a total of 36 design configurations.

For each design configuration, 100 repetitions were run as follows:

1. 50 training observations were created from a multivariate Gaussian with specified population correlation matrix and ϵ was drawn from a univariate Gaussian with the chosen σ^2 .
2. Estimators were applied to these data.
3. $N = 100$ validation observations were generated as in step 1.
4. For each estimator the average squared prediction error was computed using the estimated coefficients from the training set with the solution transformed back into the original coordinate system and unstandardized. These prediction errors were averaged over the 100 Monte Carlo repetitions.

5 Results and Conclusions

Since a number of the estimators we have discussed require a value for σ , we only discuss the situations with $p = 5, 40$. In future work, we plan to examine the $p = 100$ case in more detail. Again following Frank and Friedman (1993), we have plotted (Figure 1) the Euclidean distances of the average performance (squared prediction error) of an estimator on the validation set from the average performance of the true regression model. Lower bar heights indicate better performance.

Our results for LS, CVRR, and CVPCR are very close to those obtained by Frank and Friedman for $p = 5$ and 40 (see Figure 2 for the results disaggregated by number of variables). We have used their results to place a bar for PLS on our plots. PLS was not a part of our simulation. Our conclusions are based on Figure 1. Accordingly, the best estimator is still CVRR as in Frank and Friedman. However, HKA and HKUBA did better than CVPC and PLS. HKA and HKUBA are very simple, but ad hoc estimators and they require an estimate of σ^2 , which CVRR, CVPCR and PLS do not. For $p > n$, there is no clear way to compute HKA or HKUBA except to find σ by cross-validation. But if we are going to do that, we may as well use CVRR or PLS.

The generalized cross-validation ridge estimator, GCVRR, did not do as well as we had expected and indicates that a price must be paid for the computational simplicity.

CVPCR does better than CVTPCR and it appears to be better to use eigenvalues to order the components rather than t -values, as we had noted earlier. We also see that CPCR, which does not use cross-validation to choose the

components, does slightly better than CVPCR. The same is true (only more so) when comparing TPCR with CVTPCR.

We conclude that for $p < n$, it is feasible to find non-cross-validated ridge-type estimators that are quite competitive with PLS and not too far from CVRR. For $p \geq n$, cross-validation appears to be essential.

Acknowledgments

We would like to thank Geoffrey Lauprete for his programming and error-correcting skills and Christine Liberty for her word-processing and technical assistance.

References

- Frank, I.E. and Friedman, J.H. (1993), "A Statistical View of Some Chemometrics Regression Tools," *Technometrics*, 35, 109-148.
- Garthwaite, P.H. (1994), "An Interpretation of Partial Least Squares," *Journal of the American Statistical Association*, 89, 122-127.
- Golub, G.H., Heath, M. and Wahba, G. (1979), "Generalized Cross-validation as a Method for Choosing a Good Ridge Parameter," *Technometrics*, 21, 215-224.
- Gruber, M.H.J. (1998), *Improving Efficiency by Shrinkage*, New York: Marcel Dekker.
- Hoerl, A.E., and Kennard, R.W. (1970), "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, 8, 27-51.
- Hoerl, A.E., Kennard, R.W., and Baldwin, K.F. (1975), "Ridge Regression: Some Simulations," *Communications in Statistics: Theory and Methods*, 4(2), 105-123.
- Krzanowski, W.J. and Marriott, F.H.C. (1995), *Multivariate Analysis, Part 2, Classification, Covariance Structures and Repeated Measurements*, London: Halsted Press.
- Lawless, J.F. and Wang, P. (1976), "A Simulation Study of Ridge and Other Regression Estimators," *Communications in Statistics: Theory and Methods*, 5, 307-323.
- Rawlings, J.O. (1988,1998), *Applied Regression Analysis*, 1st ed., 2nd ed., Belmont, CA: Wadsworth.
- 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.

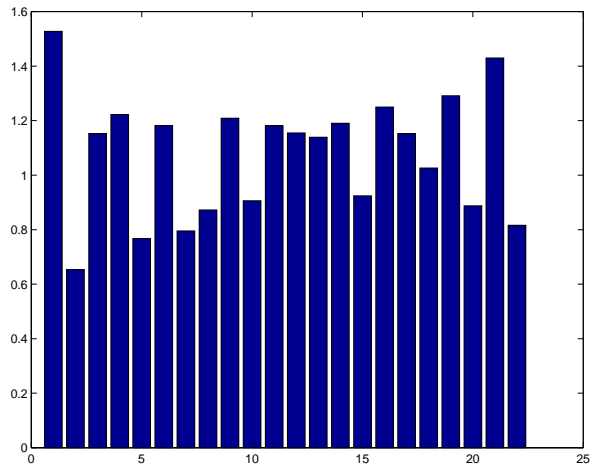


Figure 1. Simulation results for all experimental design dimensions.

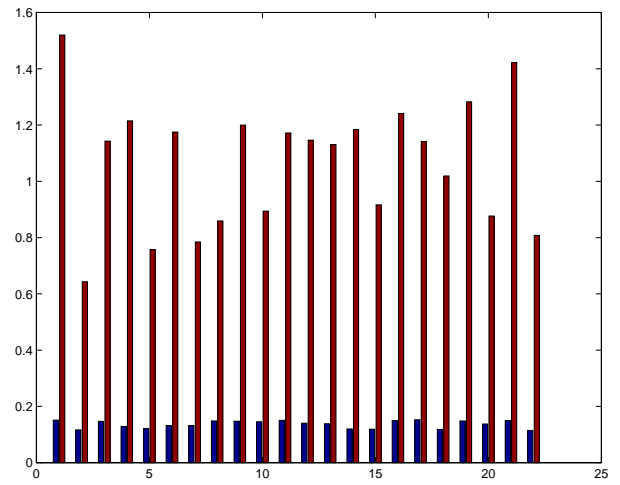


Figure 2. Simulation results disaggregated by number of variables ($p = 5$ (first bar) and 40).

Key to plots:

- | | |
|----------|------------|
| 1. LS | 12. GLW |
| 2. CVRR | 13. GLWUB |
| 3. GCVRR | 14. MSE |
| 4. HK | 15. MSPE |
| 5. HKA | 16. UBMSE |
| 6. HKUB | 17. UBMSPE |
| 7. HKUBA | 18. HKB |
| 8. CPCR | 19. LW |
| 9. TPCR | 20. CVPCR |
| 10. WPCR | 21. CVTPCR |
| 11. RPCR | 22. PLS |