Validity Concentration Formula Validation

Mary G. Lieberman **John D. Morris**
Florida Atlantic University

A formula for estimation of the validity concentration of a prediction model is specified, and its performance (ρcv and MSEcv) in tracking the accuracy of alternate non-OLS weighting methods (ridge regression, regression on principal components and equal weighting) was investigated. The performance of all alternate methods was tracked very accurately; the formula predicted performance perfectly (R2 = 1.0) for all three alternatives for both ρcv and MSEcv. The use of this validity concentration formula to help decide when to use a non-OLS method is thus deemed promising.

The purpose of this investigation was to specify and examine an objective and, hopefully useful, definition of validity concentration. Previous work has shown that the performance of various non-OLS prediction modeling techniques depend on validity concentration, and in the most recent, candidate formulas for validity concentration were proposed (Morris & Lieberman, 2016). Although the concept of validity concentration has been described in the literature, identifying a specific formula that predicts the performance of non-OLS predictor weighting techniques does not appear to have been accomplished. The goal of this study is to consider the performance of one possibility.

Theoretical Framework

Darlington (1978) posited that the relative multiple regression cross-validation prediction accuracy between OLS and alternatives, with a concentration on ridge regression, is a function of R^2, N, p, and validity concentration, where R^2 represents the squared sample multiple correlation, N is sample size, and p is the number of predictor variables. In Darlington's formulation, validity concentration was used to describe a data condition in which the principal components of the predictor variables' intercorrelation matrix with large eigenvalues also have large correlations with the criterion. Thus, validity concentration requires at least a modicum of predictor variable collinearity (large predictor eigenvalues); but collinearity is only necessary, not sufficient, for validity concentration. Morris (1982) examined the prediction performance of the specific version of ridge regression recommended by Darlington (RIDGM, Dempster, Shatzoff, & Wermuth, 1977) with the same data structures on which the technique's superiority was posited. A synopsis of those results were that OLS was superior at smaller levels of validity concentration and as validity concentration increased, alternatives to OLS became superior. So, as Darlington specified, Ridge became better than OLS with larger validity concentration, but also, in those same conditions, other methods exceeded Ridge.

In addition, it has been shown that in the case of classification (Morris & Huberty, 1987; Morris, & Lieberman, 2012), prediction accuracy follows a similar pattern. As well, it was shown that OLS prediction performance was unrelated to multicollinearity, regardless of whether validity concentration obtained (Morris & Lieberman, 2015). The difference that accrues between OLS and alternative methods is not due to a loss in predictive performance of OLS as multicollinearity increases (as stated, performance is flat in respect to multicollinearity), rather it is due to the enhancement in performance of alternative methods afforded by increasing validity concentration (given, of course, the requisite increase in multicollinearity).

Method

As specified, the notion of validity concentration has been vaguely described. In a comparison of the prediction performance of OLS and alternatives, Darlington (1978) posed the term "validity concentration" and manipulated it in the following way. The data situation was such that the number of predictors (p) was 10. Requisite predictor variable multicollinearity was manipulated such that principal component eigenvalues of the predictor variable intercorrelation matrix decreased by a constant ratio (λ_r =.50, .65, .80 and .95), therein creating decreasing multicollinearity with increasing λ_r. In turn, validity concentration was manipulated such that corresponding principal components manifested squared correlations with the criterion that were proportional to varying powers of the eigenvalues; those powers were .1, .5, 1, 2, 4, and 10, with the resulting squared component validities necessarily summing in each case to the prescribed R^2 of .25. Thus, therein, creating increasing validity concentration, within λ_r, as this power increased. However, as stated, multicollinearity limits the degree of validity concentration possible, thus the validity concentration was, in each case, capped by λ_r.

The purpose herein is to consider a possible formula for validity concentration and to examine how well it performs in detecting the data conditions under which some alternative methods outperform OLS.

A potentially appealing index of validity concentration, might be:

$$VC = (\Sigma\lambda_i\rho_i^2/R^2 - 1)/(p - 1), \quad (1)$$

where λ_i and ρ_i^2 are the ith eigenvalue of a principal component of the predictor variable intercorrelation matrix and its corresponding squared component validity, R^2 is the sample squared multiple correlation, and p is the number of predictor variables. As the predictor variable intercorrelation matrix is full rank (else OLS multiple correlation would be impossible), and as the component validity is squared, both λ_i and ρ_i^2 are necessarily positive. VC can be negative under the unusual circumstance of large eigenvalues being systematically associated with smaller component validities. However, although it is certainly possible to create such a situation "in the lab," with components manifesting more variance having less covariance shared with the criterion, this is deemed very unlikely to be the case with real data. Outside of this possibility, VC is nicely bounded [0,1]. These limits seem consistent with the aforementioned notion of validity concentration.

Considering the operating characteristics of the formula, as λ_1 approaches p, representing perfect collinearity in the first component (of course to be considered here only as an asymptotic theoretical limit, as OLS regression would be impossible in that case due to the resultant singular intercorrelation matrix), and ρ_1^2 approaches R^2 (given that λ_1=p, representing maximum validity concentration), the numerator becomes p-1, and thus VC=1. On the other hand, if all λ_i=1, representing minimum multicollinearity – indeed, orthogonality -- the numerator becomes zero, thus VC = 0, regardless of the ρ_i^2s. If validity is equally distributed across components such that all $\rho_i^2=R^2/p$, the same obtains with VC=0, regardless of the λ_is. This can be interpreted in line with former arguments; validity concentration can't exist without a modicum of collinearity, but collinearity, no matter how great, can't manifest validity concentration without the association of larger ρ_i^2s with larger λ_is.

To allow maximum validity concentration range therefore requires multicollinearity, thus a λ_r of .30 was used herein to examine the performance of VC. To provide context, one further comment (and the reason for its selection) about the .30 λ_r condition is needed. Although originally posited as a test of the ability of digital computers to accomplish the necessary inversion of a near singular matrix for regression (Longley, 1967), the infamous "Longley Data" has often been used as a reference point for **very** extreme multicollinearity (VIFs from 4 to 1789). With VIFs of 340 to 2000, the .30 λ_r condition manifested even greater multicollinearity than the Longley data.

Validity concentration was manipulated as in former studies with powers of the λ of .1, .5, 1, 2, 4, and 10, but with additional powers of 0, .2, .3, .4, .7, and 1.5 added to allow more precise examination of the functional relationship between the VC index and model accuracy. [Note that these powers are, and have been, used to **manipulate** validity concentration; they are not a **measure** of validity concentration as is the VC index.] Consistent with former studies mentioned, N and R^2 were set at 40 and .25.

The alternatives to OLS considered herein were bounded ridge regression, regression on principal components with dimensionality decided by parallel analysis, and regression on equal weights; all executed as in the aforementioned Morris and Lieberman (2015) study.

A population of 10,000 subjects was created (Morris, 1975; 1982) that manifested each condition. Samples of 40 subjects were selected with 1000 replicates. Prediction models as specified were created from the sample and cross-validated by predicting the criterion in the population for all 10,000 Ss. A Fortran 90 computer program compiled by Intel Parallel Studio XE 2016 was used to accomplish all simulations.

Results

The relationships between the alternative methods' cross-validated prediction performance (correlation with the criterion, ρ_{cv},, and MSE_{cv}) and VC are illustrated in Figures 1 and 2, respectively. OLS is also included as a reference, but, as mentioned, its performance is flat in respect to VC; OLS performance is totally insensitive to multicollinearity or potential attendant validity concentration.

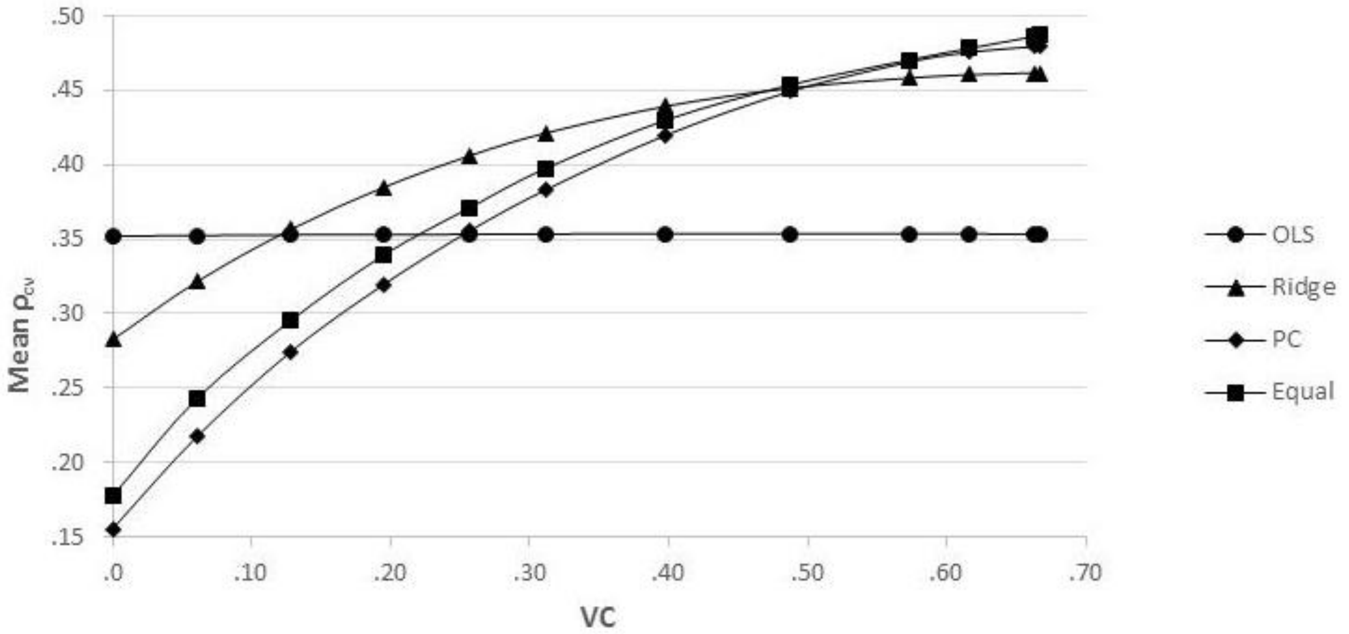

Figure 1. Methods' ρ_{cv} as a function of VC.

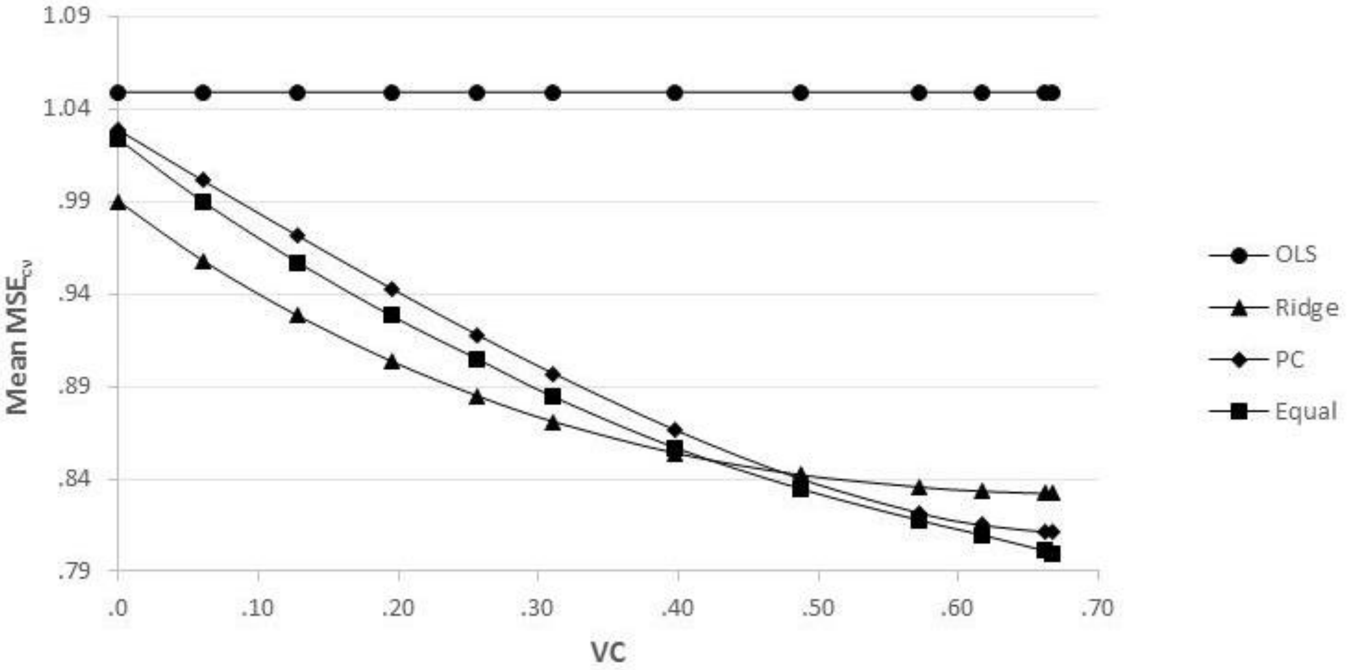

Figure 2. Methods' MSE_{cv} as a function of VC.

The question then, is whether the VC index provides an accurate functional relationship with the performance of the alternative methods. As the alternative methods capitalize on validity concentration, for the VC index to be of predictive use, as it increases, one would hope for a clear functional relationship with non-OLS model performance. For each method, a simple, and smooth, monotonically increasing, or decreasing, function resulted for $\rho_{cv,}$, and MSE_{cv}, respectively. The relationship was clearly curvilinear; increasing VC increases the performance of the alternative methods, but the degree of that increase lessens as VC increases. Therein, Ridge capitalizes on increasing VC first and remains superior to the other alternative methods (PC and Equal) over the greatest range of VC. As VC increases even more, PC and Equal exceed the accuracy of Ridge. This is also consistent with the results from previous studies mentioned.

To be specific, the question is predictive. How well can the VC index predict the performance of an alternative to OLS that takes advantage of validity concentration? As has been pointed out, visual inspection of curve fit would lead one to judge fit as excellent for all three alternative methods. In addition, for both ρ_{cv} and MSE_{cv}, indices of accuracy for a cubic fit on VC was used, and from that, a perfect R^2=1.0 resulted for all three methods. Thus the cubic model provides perfect prediction of both ρ_{cv} and MSE_{cv} performance for all three alternatives to OLS. As they capitalize on validity concentration to differing degrees, they differ in their coefficients (see Table 1).

Table 1. Betas for Cubic Fit of Each Non-OLS Method (R^2 = 1.0 for All Models)

	Ridge		PC		Equal	
	ρ_{CV}	MSE_{CV}	ρ_{CV}	MSE_{CV}	ρ_{CV}	MSE_{CV}
VC	2.670	-2.470	2.207	-1.336	2.458	-1.772
VC^2	-2.485	2.055	-1.659	-0.130	-2.333	1.013
VC^3	0.733	-0.509	0.369	0.510	0.825	-0.221

VC appears to be a potentially useful index of validity concentration that may help in consideration of when one might employ such non-OLS prediction algorithms. More examination is, of course, needed. Included herein are population VC values; its sampling distribution is unknown. As well, although performance of the formula with real data sets has been accomplished and seems promising, more data sets, with a wider distribution of relevant characteristics are sought, particularly from the *GLMJ* readership.

References

Darlington, R. B. (1978). Reduced variance regression. *Psychological Bulletin*, *85*, 1238-1255.

Dempster, A. P., Schatzoff, M., & Wermuth, N. (1977). A simulation study of alternatives to ordinary least squares. *Journal of the American Statistical Association*, *72*, 77-91.

Lieberman, M. G., & Morris, J. D. (2016, April). *Validity concentration formula validation*. Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C.

Longley, J. W. (1967). An appraisal of least-squares programs from the point of view of the user. *Journal of the American Statistical Association, 62*, 819–841.

Morris, J. D. (1975). A computer program to create a population with any desired centroid and covariance matrix. *Educational and Psychological Measurement*, *35*, 707-710.

Morris, J. D. (1982). Ridge regression and some alternative weighting techniques: A comment on Darlington. *Psychological Bulletin*, *91*, 203-210.

Morris, J. D., & Huberty, C. J. (1987). Selecting a two-group classification weighting algorithm. *Multivariate Behavioral Research*, *22*, 211-232.

Morris, J. D., & Lieberman, M. G. (2012). Selecting a two-group classification weighting algorithm: Take two. *Multiple Linear Regression Viewpoints*, *38*, 34-41.

Morris, J. D., & Lieberman, M. G. (2015). Prediction, explanation, multicollinearity and validity concentration in multiple regression. *General Linear Model Journal*, *41*, 29-35.

Send correspondence to: Mary G. Lieberman
Florida Atlantic University
Email: mlieberm@fau.edu