The Precise Effect of Multicollinearity on Classification Prediction

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

The results of Morris and Lieberman (2012) were extended to include more extreme multicollinearity conditions, with specific attention paid to the effect of a wide range of multicollinearity levels on Ordinary Least Squares and Logistic Regression cross-validated classification accuracy, and the attendant increase in prediction accuracy afforded by Ridge Regression and Principal Components models with increasing validity concentration. Even very extreme multicollinearity did not affect OLS or LR prediction accuracy, but, given attendant validity concentration, did enable Ridge Regression and Principal Components to exceed the accuracy of these typically used classification techniques. Clarification regarding how these results bear on admonitions regarding multicollinearity was tendered.

This investigation extends Morris and Lieberman (2012), which contrasted the accuracy of six algorithms for classifying subjects into one of two groups. Although the 2012 study considered a wide variety of alternate prediction methods, degrees of group separation, and validity concentration levels, only one of the levels of multicollinearity included would normally be deemed serious (Variance Inflation Factor (VIF) > 10). This study extends multicollinearity conditions considerably, paying particular attention to the effect of multicollinearity on the traditionally used Ordinary Least Squares (OLS) and Logistic Regression (LR) models, and the enabling effect of multicollinearity in respect to validity concentration for alternate methods.

Darlington (1978) posited that the relative performance of OLS versus alternative methods' regression cross-validation accuracy is a function of R^2, N, and validity concentration, where R^2 represents the sample squared multiple correlation and N the sample size. Darlington described validity concentration as a data condition in which the principal components of the predictor variable intercorrelation matrix, with large eigenvalues, also have large correlations with the criterion. Thus, validity concentration inherently depends on predictor variable collinearity; a modicum of predictor variable collinearity is necessary to result in large eigenvalues. But, collinearity is only necessary, not sufficient, for validity concentration. In respect to regression, Morris (1982) examined the performance of a variety of prediction models with the data structures posed by Darlington. Most pertinent to this study, Morris and Huberty (1987) examined a subset of the methods considered in Darlington and Morris (1982) [OLS, Ridge Regression (RR), and Principle Components (PC)] in the context of two-group classification accuracy (rather than regression) using the same simulated data conditions.

Morris and Lieberman (2012) used the same data conditions as in Morris and Huberty (1987), but included the Pruzek and Frederick (1978) and Equal Weighting techniques. In addition, LR was included. A synopsis of those results were that OLS and LR were superior at smaller levels of validity concentration (disregarding occasions in which LR estimates failed to converge due to maximum likelihood failure – a necessity with complete group separation in the sample) and as validity concentration increased, the alternate methods became superior. As in Darlington (1978), and all subsequently cited studies, multicollinearity was manipulated by creating data with varying predictor variable eigenvalue ratios (λ_r) of .95, .80, .65, and .50. These were simply specified as the proportion of decrease in the principal component eigenvalues in a 10 variable prediction problem. Thus, as the λ_r becomes smaller, multicollinearity of the predictors becomes larger. Therefore, the .50 λ_r condition represented the highest level of multicollinearity afforded in all of these studies. VIFs using the .50 λ_r condition ranged from about 12 to 39; thus, representing what might normally be judged a multicollinear predictor variable set. However, none of the other λ_r conditions manifested VIFs that would usually be considered multicollinear. As well, the .50 λ_r condition does not manifest the higher levels of multicollinearity often seen in real data sets. Thus, part of the goal of this study was to examine classification accuracy under more extreme multicollinearity conditions.

Method

The point of this study was to extend results regarding the contrast between the standard classification methods (OLS and LR) and alternatives. The purpose was not an examination of the relative performance of OLS versus LR. The commonly purported advantage of LR regards immunity to the distributional requirements of OLS; thus, as distribution shape was not manipulated in this study, such comparisons are judged inappropriate. The purpose was also not to identify the best alternate method, rather to consider accuracy trends of OLS and LR versus some typically available alternatives. Therefore, other than the addition of LR, the same methods as included in the previously mentioned Morris and Huberty (1987) original presentation of alternatives to OLS prediction in the classification context were used herein. The methods used were:

- OLS. Fisher's LDF.
- Ridge. Empirical ridge regression weights calculated using the biasing parameter, *k*, due to Lawless and Wang (1976).
- PC. Prediction from principal components with the number of components chosen by parallel analysis. This is a distinction from former studies in which the relatively naïve $\lambda > 1.0$ rule was used to decide dimensionality.
- LR: The logistic regression maximum likelihood solution was calculated with iteration using the Newton-Raphson method. Mimicking SPSS v. 22, iteration was halted when all weights changed by no more than .001. Failure in solution generation was captured as caused by 1) complete separation, 2) a singular weight covariance estimate occurrence at an iteration, or 3) non-convergence of weights that were increasing to infinity. The same tolerance for extreme probabilities (10^8) as used in SPSS v. 22 was employed.

All data conditions were such that there were 10 predictor variables. For Darlington's condition of a constant proportion eigenvalue decrease, it is obvious that if the first eigenvalue can be calculated, all remaining eigenvalues are evident. The formula for calculating the first eigenvalue is **$\lambda_1=p/(1 + \sum_{j=1}^{q} \lambda_r^{j})$**, where p is the number of components (or variables), and q = p-1. Increased levels of collinearity were created such that λ_rs of .30 and .40 were used in addition to the previously mentioned .50 and .65 levels. This resulted in four conditions in which the VIFs were what would normally be considered problematic (.50 λ_r: VIFs ranged from about 12 to 39), very problematic (.40 λ_r: VIFs ranged from about 49 to 232), and what might be considered by many to be tragically problematic (.30 λ_r: VIFs ranged from about 340 to 2000). In addition, one condition in which multicollinearity would be considered benign (.65 λ_r: VIFs ranged from about 3 to 6) was included. 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 manifests greater multicollinear than the Longley data. It would be difficult to argue that an extreme range of multicollinearity has not been covered herein.

Validity concentration was manipulated in exactly the same way, and with the same levels, as in the former cited studies. Six levels of validity concentration were created such that the component validities were proportional to a power of the eigenvalues; powers of .1, .5, 1, 2, 4, and 10 were used. Thus, as the squared component validities had to sum to the desired multiple correlation, they were also uniquely determined.

Procedure

As in Morris and Huberty (1987), these 24 conditions (4 λ_r by 6 validity concentration conditions) were expanded in a fully crossed design to include two population multiple correlations (ρ^2) of .25 and .75. A population of 10,000 subjects was created to manifest each of the desired sets of collinearity, validity concentration, and multiple correlation characteristics. Manipulating sample size was done by simply selecting 40, or alternatively 100, subjects in a sample (with replacement). The algorithm used to create the population covariance matrices representing the respective eigenstructure and validity concentration is described in Morris (1982), and the algorithm for creating the population of subjects manifesting each desired covariance matrices is described in Morris (1975).

To transform each of these data structures into a two-group classification problem, each population was dichotomized at the median criterion score and the prediction weights from randomly selected samples of the desired size were calculated to predict the dichotomous criterion by each of the prediction

Table 1. OLS and LR Performance by Multicollinearity Condition

		Mean Performance			
λ_r	VIF Range	$\rho^2 = .25$		$\rho^2 = .75$	
	(rounded)	OLS	LR	OLS	LR
.30	340 to 2000	.6136	.6127	.7841	.7808
.40	49 to 232	.6102	.6103	.7820	.7807
.50	12 to 39	.6082	.6077	.7821	.7769
.65	3 to 6	.6121	.6115	.7857	.7813
.80	1 to 2	.6085	.6083	.7809	.7767
.95	< 1	.6064	.6060	.7761	.7723
F(5,66) =		0.22	0.19	0.26	0.19

methods. The sample weights were then cross-validated by using them to classify all 10,000 population subjects. The total number of correct classifications was used to compare the relative accuracies of the methods. Toward greater stability, this procedure was repeated 1000 times within each data condition (rather than 100 replicates as in former studies) with the mean classification accuracy representing the accuracy of each method. The random normal deviates required were created by the "Rectangle-Wedge-Tail" method (Marsaglia, MacLauren, & Bray, 1964), with the required uniform random numbers generated by the "shuffling" Algorithm M recommended by Knuth (1969, p. 30). Dolker and Halperin (1982) found this combination to perform most satisfactorily in a comparison of several methods of creating random normal deviates. A Fortran 90 computer program compiled by Intel Parallel Studio XE 2013 was used for accomplishing all simulations.

Results and Discussion

In consideration of the effect of multicollinearity on traditional classification methods, Table 1 shows the performance of OLS and LR methods for the .25 and .75 ρ^2 conditions as a function of λ_r, collapsed across all other conditions. Only the previously mentioned .30, .40, and .65 λ_r conditions were considered henceforth in this paper, but in Table 1, the nearly orthogonal additional λ_r conditions of .80 and .95 were presented to capture the effect of a very broad range of collinearity conditions on prediction accuracy for traditional classification algorithms. Additionally an F-ratio from a one-way analysis of variance comparing the means across the six multicollinearity levels is included. One can see that the answer is very simple; multicollinearity, from non-existent to very extreme, has absolutely no effect on the cross-validated classification accuracy of OLS or LR for either group separation level (ρ^2). One might note that the F-ratios never even approached 1. Moreover, if one considers only two significant digits, variation in accuracy across multicollinearity levels (within a ρ^2 level) is a maximum of .01 for both OLS and LR.

This is not to say that multicollinearity is unimportant; only that it is not important to cross-validated prediction accuracy. Philosophical precision regarding the goal of the research is necessary. The distinction between the goal of prediction, wherein interest is in creating a model that is maximally accurate versus that of explanation, in which judgments regarding the explanatory power of variables within a model is the interest, has been nicely detailed elsewhere (Kerlinger, 1973, p. 9-10; Kerlinger & Pedhazur, 1973, p. 48-49; Huberty, 2003). If prediction accuracy is the goal, then multicollinearity is not relevant. If, on the other hand, attention does not regard model prediction accuracy, but allocation of variable importance within the model, then, of course, multicollinearity tends to confound such judgment. If interest is in both goals, the researcher may need to contemplate their relative importance. However, it seems that what appear to be blanket warnings regarding multicollinearity in classification problems should at least be more precisely expanded in consideration of these alternate modeling goals (Hosmer & Lemeshow, 2000, p. 140-141; Schaefer, 1986). Herein, interest is in classification; however, the same comment can certainly be made regarding regression, or other prediction methods.

Tables 2 and 3 include the classification performance for all methods at each λ_r, validity concentration (power), and sample size for the .25 and .75 ρ^2 conditions, respectively. Rounded VIF ranges are also

Table 2. Mean Proportion Correctly Classified for ρ2 = .25 (Mean Mahalanobis D2 = .79)

		Method							
λ_r	Power	N = 40				N = 100			
[VIF Range]		OLS	Ridge	PC	LR/#*	OLS	Ridge	PC	LR/#*
.30	.1	**.6002**	.5728	.5534	**.5980**/969	**.6361**	.6024	.5724	**.6358**/1000
[340 to 2000]	.5	.5989	**.6186**	.6065	.5968/973	.6339	**.6433**	.6260	.6337/1000
	1.	.5943	**.6288**	.6265	.5928/972	.6293	**.6504**	.6452	.6294/1000
	2.	.5935	.6344	**.6388**	.5920/977	.6291	.6540	**.6557**	.6292/1000
	4.	.5944	.6359	**.6414**	.5930/976	.6293	.6543	**.6568**	.6294/1000
	10.	.5949	.6367	**.6422**	.5933/978	.6293	.6544	**.6568**	.6293/1000
.40	.1	**.5926**	.5696	.5463	**.5911**/980	**.6297**	.5999	.5634	**.6294**/1000
[49 to 232]	.5	.6003	**.6145**	.5980	.5984/966	.6364	**.6422**	.6171	.6359/1000
	1.	.5987	**.6296**	.6252	.5967/974	.6344	**.6524**	.6426	.6342/1000
	2.	.5934	.6300	**.6349**	.5917/970	.6291	.6511	**.6528**	.6291/1000
	4.	.5945	.6331	**.6403**	.5930/979	.6303	.6530	**.6572**	.6304/1000
	10.	.5943	.6331	**.6406**	.5929/979	.6296	.6525	**.6569**	.6296/1000
.50	.1	**.5936**	.5730	.5469	**.5924**/984	**.6271**	.6031	.5666	**.6272**/1000
[12 to 39]	.5	.5929	**.6033**	.5868	.5911/981	.6259	**.6302**	.6088	.6259/1000
	1.	.5933	**.6186**	.6134	.5914/977	.6271	**.6428**	.6341	.6270/1000
	2.	.5959	.6285	**.6342**	.5939/971	.6309	.6503	**.6521**	.6307/1000
	4.	.5941	.6289	**.6383**	.5923/973	.6301	.6504	**.6551**	.6300/1000
	10.	.5934	.6279	**.6380**	.5916/981	.6286	.6491	**.6541**	.6286/1000
.65	.1	**.5993**	.5859	.5562	**.5970**/967	**.6351**	.6216	.5779	**.6350**/1000
[3 to 6]	.5	**.5979**	.5997	.5798	**.5956**/970	**.6329**	.6306	.6033	**.6327**/1000
	1.	.5963	**.6101**	.6010	.5942/971	.6309	**.6383**	.6247	.6308/1000
	2.	.5939	.6163	**.6220**	.5921/976	.6288	.6416	**.6431**	.6286/1000
	4.	.5957	.6217	**.6332**	.5939/971	.6310	.6458	**.6526**	.6308/1000
	10.	.5955	.6231	**.6356**	.5933/974	.6315	.6470	**.6540**	.6314/1000

Note: The best performing method or set of methods is in bold (p<.01).

* Number of samples for which LR calculable.

included. As λ_r increases, multicollinearity becomes less; in turn, within each λ_r, validity concentration becomes larger as power increases. Thus, for example, the λ_r of .30 and power of .1 offers the highest multicollinearity and lowest validity concentration (within that λ_r condition).

An overall Hotelling T^2 test, and then post hoc comparisons, was used to contrast the methods' classification hit-rates as in Morris and Huberty (1987). As interest was not in differentiating between OLS and LR, post hoc linear contrasts were between OLS and LR together and; alternatively, Ridge and PC. In line with Morrison's suggestion (1976, p. 148), the means, as specified above, were ordered and pairwise post hoc contrasts were used to contrast larger means with smaller means until a significant ($p <$ .01) difference was found, thus delineating best methods.

The results were so clear that both tables can be discussed together. As would be expected, trends regarding validity concentration were parallel to those found in Morris and Huberty (1987) and Morris and Lieberman (2012). One will note many samples, particularly of size 40, and more so for the higher

Table 3. Mean Proportion Correctly Classified for ρ2 = .75 (Mean Mahalanobis D2 = 3.69)

		Method							
λ_r	Power	N = 40				N = 100			
[VIF Range]		OLS	Ridge	PC	LR/#*	OLS	Ridge	PC	LR/#*
.30	.1	**.7681**	.6958	.6290	**.7586**/431	**.8079**	.7317	.6481	**.8081**/999
[340 to 2000]	.5	.7656	**.7836**	.7327	.7573/491	**.8050**	**.8056**	.7475	**.8057**/1000
	1.	.7637	**.8061**	.7876	.7569/485	.8047	**.8201**	.8002	.8054/1000
	2.	.7639	.8103	**.8128**	.7564/466	.8032	.8214	**.8223**	.8045/1000
	4.	.7622	.8101	**.8167**	.7549/465	.8018	.8212	**.8250**	.8037/1000
	10.	.7612	.8097	**.8168**	.7549/482	.8016	.8214	**.8254**	.8035/1000
.40	.1	**.7683**	.7089	.6289	**.7609**/449	**.8087**	.7475	.6463	**.8094**/999
[49 to 232]	.5	.7680	**.7786**	.7170	.7584/452	**.8082**	.8043	.7301	**.8084**/999
	1.	.7647	**.8016**	.7748	.7566/481	.8060	**.8184**	.7852	.8067/1000
	2.	.7655	**.8083**	**.8098**	.7593/480	.8054	**.8211**	.8187	.8062/1000
	4.	.7622	.8059	**.8155**	.7546/473	.8024	.8187	**.8244**	.8038/1000
	10.	.7632	.8073	**.8176**	.7561/452	.8036	.8206	**.8268**	.8052/999
.50	.1	**.7574**	.7093	.6201	**.7520**/486	**.7997**	.7579	.6464	**.8015**/999
[12 to 39]	.5	.7647	**.7721**	.7079	.7574/471	**.8063**	.8026	.7298	**.8072**/999
	1.	.7669	**.7975**	.7640	.7580/493	.8084	**.8184**	.7828	.8086/999
	2.	.7649	**.8030**	**.8023**	.7561/500	.8054	**.8191**	.8155	.8060/1000
	4.	.7658	.8047	**.8165**	.7598/469	.8062	.8199	**.8260**	.8072/1000
	10.	.7625	.8013	**.8138**	.7543/484	.8022	.8161	**.8227**	.8032/998
.65	.1	**.7660**	.7407	.6295	**.7567**/450	**.8052**	.7916	.6559	**.8060**/999
[3 to 6]	.5	**.7654**	**.7631**	.6830	**.7568**/463	**.8041**	.8008	.7089	**.8051**/1000
	1.	.7654	**.7809**	.7360	.7567/471	.8040	**.8080**	.7585	.8047/998
	2.	.7702	**.7964**	.7892	.7615/499	.8090	**.8171**	.8079	.8095/1000
	4.	.7677	.7979	**.8116**	.7577/474	.8086	.8172	**.8254**	.8087/999
	10.	.7664	.7968	**.8146**	.7582/466	.8071	.8162	**.8258**	.8075/999

Note: The best performing method or set of methods is in bold (p<.01).

* Number of samples for which LR calculable.

group separation, for which LR failed to converge. This is due to complete, or quasi-complete, group separation in the sample and is discussed in Morris and Lieberman (2012). OLS and LR were superior at smaller levels of validity concentration. As validity concentration increased within λ_r level, Ridge became superior, and then at even higher validity concentration, PC became superior. The larger sample size (N = 100) and group separation (ρ^2 = .75) tended to retard this trend to a minor degree.

Further attention to the influence of multicollinearity and, in turn, validity concentration is needed. It is clear that multicollinearity, by itself, had no effect on cross-validated prediction accuracy of OLS and LR methods, and it only had an indirect effect (through “capping” the possible validity concentration) on the performance of the alternate methods. Across all multicollinearity levels, sample size, and group separation conditions, at the lowest level of validity concentration (power = .1), one can see that OLS and LR were clearly superior to the alternate methods. It is only upon validity concentration increasing that the alternative methods fair better than OLS and LR. One notes that the aforementioned advantage of the alternate methods induced by increasing validity concentration is lessened as multicollinearity decreases

(λ_r increases). This is because multicollinearity is necessary to obtain validity concentration; it is not possible for large eigenvalues to align with large component validities if there are no large eigenvalues.

The important lesson herein is that it is not a decline in accuracy of OLS or LR with increasing multicollinearity or validity concentration that advantages these alternative methods; indeed OLS and LR are insensitive to both multicollinearity and validity concentration. The superior performance of the alternative methods over OLS and LR arises from their ability to take advantage of validity concentration and; thus, given its presence, exceed the accuracy of OLS and LR.

References

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

Dolker, M., & Halperin, S. (1982). Comparing inverse, polar, and rectangle-wedge-tail FORTRAN routines for pseudo-random normal number generation. *Educational and Psychological Measurement*, *42*, 223-236.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York: Wiley.

Huberty, C. J (2003). Multiple correlation versus multiple regression. *Educational and Psychological Measurement*, *63*, 271-278.

Kerlinger, F. N. (1973). *Foundations of behavioral research* (2nd ed.). New York: Holt.

Kerlinger, F. N., & Pedhazur, E. J. (1973). Multiple regression in behavioral research. New York: Holt.

Knuth, D. E. (1969). *The art of computer programming* (Vol. 2: Seminumerical algorithms). Reading, MA: Addison-Wesley.

Lawless, J. F., & Wang, P. (1976). A simulation study of ridge and other regression estimators. *Communications in Statistics, A5*, 307-323.

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.

Marsaglia, G., MacLaren, D., & Bray, T. A. (1964). A fast procedure for generating random normal variables. *Communications of the ACM*, *7*, 4-10.

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.*

Morrison, D. F. (1976). *Multivariate statistical methods*. New York: McGraw-Hill.

Pruzek, R. M., & Frederick, B. C. (1978). Weighting predictors in linear models: Alternatives to least squares and limitations of equal weights. *Psychological Bulletin*, *85*, 2, 254-266.

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