Session 3

Variable Selection/Reduction Approaches to Capture Suppressor Variables

A. Importance of suppressor variables
B. Example: Simulated data with many true predictors
    1. Linear Regression with continuous dependent variable
    2. Logistic Regression/LDA with a dichotomous dependent variable
C. Failure of common prescreening methods to capture suppressor variables
D. Coefficient Path Plots

When one or more predictors are irrelevant, excluding them from a regression can improve the predictive performance of the model. (See Fig. 2B on page 7 for an example where predictor X_3 is irrelevant.) The most widely used method for variable selection in regression, is *stepwise* regression. While Session 1 listed several problems with stepwise regression, it is still currently the most widely used approach because it is readily available in the major statistical software packages.

In Exercise B2, we will compare the performance of stepwise regression with CCR for simulated data containing a number of irrelevant variables, where the valid variables contain at least one important *suppressor* variable. Suppressor variables are uncorrelated with the dependent variable, but improve predictive performance by enhancing the effects of one or more predictors included in the model. (See Fig. 2D on page 7 for an example where predictor X_5 is a suppressor, enhancing the performance of predictor X_1.) As we will see, suppressor variables are typically among the most important predictors to include in a model. However, finding suppressors is challenging because like irrelevant variables, they are uncorrelated with the dependent variable.

In our simulated data example, we will see that how well stepwise regression performs depends largely on whether or not it is successful in including the important suppressor variable in the model. When it is successful, its performance is comparable to that of CCR. When it is not successful, its performance lags behind CCR. We will see that the CCR algorithm is very effective in capturing the effects of suppressor variables in a model.

Part A below formally introduces suppressor variables. In Part B we will explore examples with simulated data separately for cases with a continuous and a dichotomous dependent variable.

A. Importance of Suppressor Variables

The most important predictor in a regression model may be a suppressor variable which does not predict the outcome variable directly but improves the overall prediction by enhancing the effects of other predictors in the model. Suppressor variables were first introduced by Horst (1941).

A prime example of suppressor variables was provided by Horst (1966) regarding the selection of pilots during World War II. To reduce the high cost and time required to select good candidates to train, the military decided to test applicants' spatial ability via written tests. Verbal ability was found to have a near zero correlation with successful training, but was highly correlated with the spatial ability test score. This positive correlation resulted simply because verbal ability was necessary to read and comprehend the written tests measuring spatial ability. Since the questions on the spatial ability test require both understanding the question (i.e., verbal ability) and visualizing the situations (i.e., spatial ability), the spatial ability test score is confounded with verbal ability.

Surprisingly, when verbal ability scores were added as an additional predictor to the regression equation, Horst found that the total R^2 increased significantly, i.e. it improved the predictive value of the spatial scores. Horst (1966) explained that this was a result of the verbal ability predictor suppressing irrelevant portions of the other predictor. That is, when combined with the spatial ability score, the verbal ability predictor removed those irrelevant parts of the spatial ability score that were not associated with pilot training, thus boosting its predictive performance. Including a test of verbal ability was found to be essential in predicting pilot training success even though this predictor alone was not correlated with the dependent variable. See Section A2 on page 10 for a path diagram specification of the model.

The scatterplot in Figure 1 provides a graphical illustration of the effects of suppression. Verbal ability is plotted on the horizontal axis, and Technical ability on the vertical axis. Persons completing the training successfully are represented by ‘X’, while those failing to complete the training are denoted by ‘O’ symbols. Confidence ellipses for both groups are added to the plot. Note the following:

1. The variables plotted have a high positive correlation for both the successful and unsuccessful training groups. This is indicated by the positive tilt in both ellipses.

2. Spatial ability alone is predictive of successful training. This is indicated by the fact that successfully trained persons tend to lie above the horizontal line at the mean spatial ability score (Fig 1A) while unsuccessfully trained persons tend to fall below this line.
3. Verbal ability alone is not predictive of successful training. This is indicated by the fact that approximately the same proportion (50%) of successfully trained persons lie to the right of the vertical line at the mean Verbal ability score as to the left of the line (Fig 1B).
4. The best prediction of those who would successfully complete the training is provided by a model based on both predictors (Fig 1C).

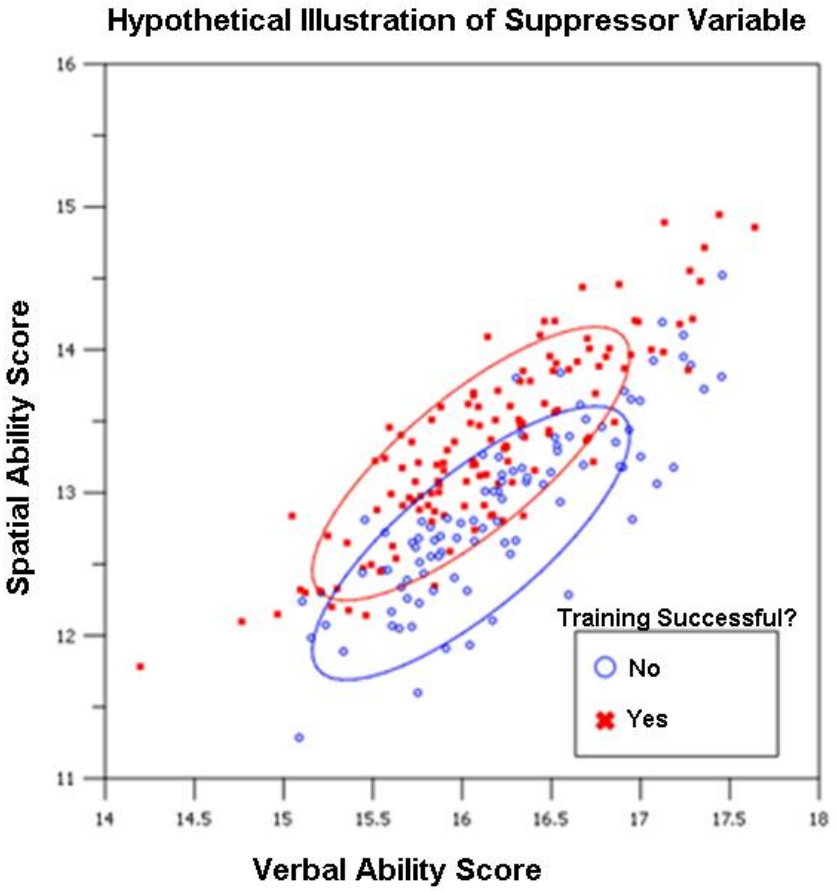


Fig. 1.

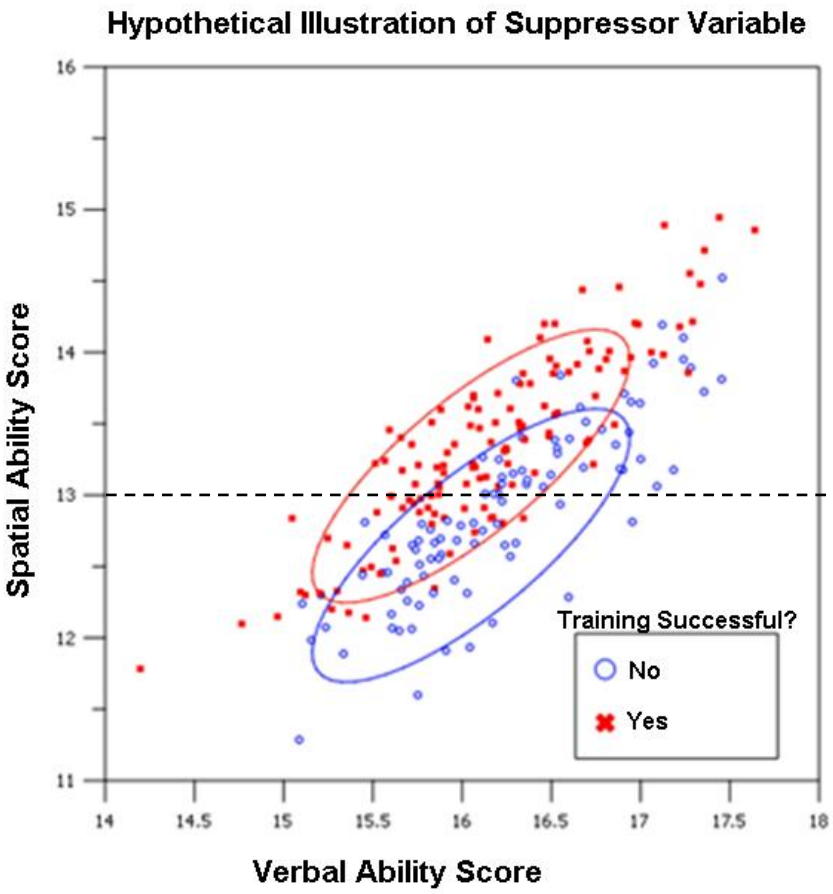


Fig. 1A.

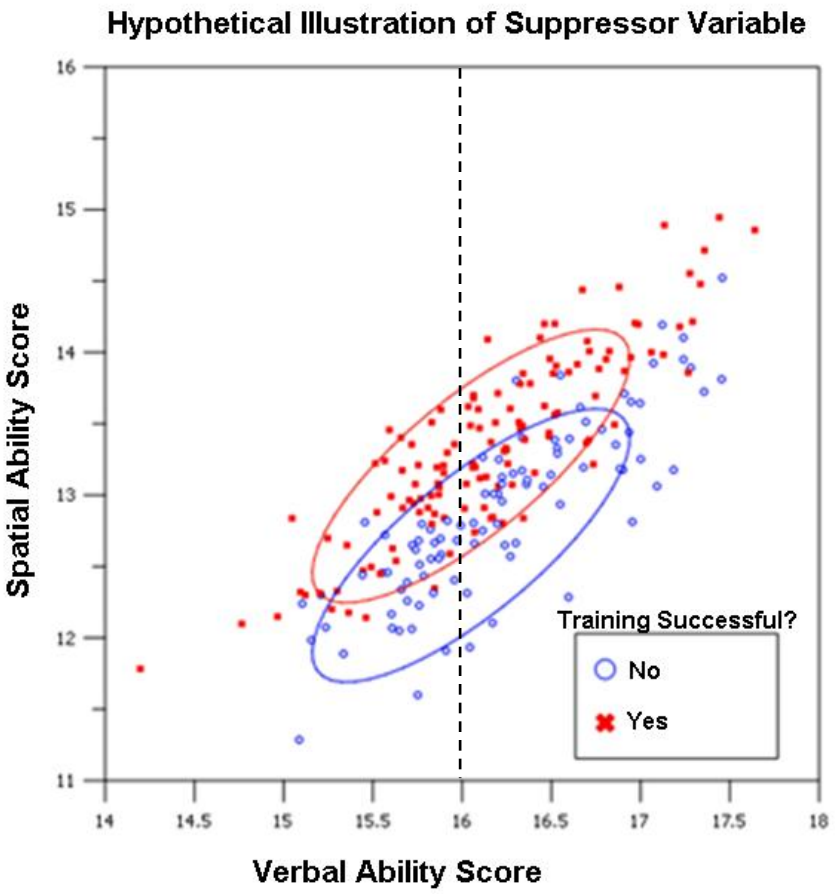


Fig. 1B.

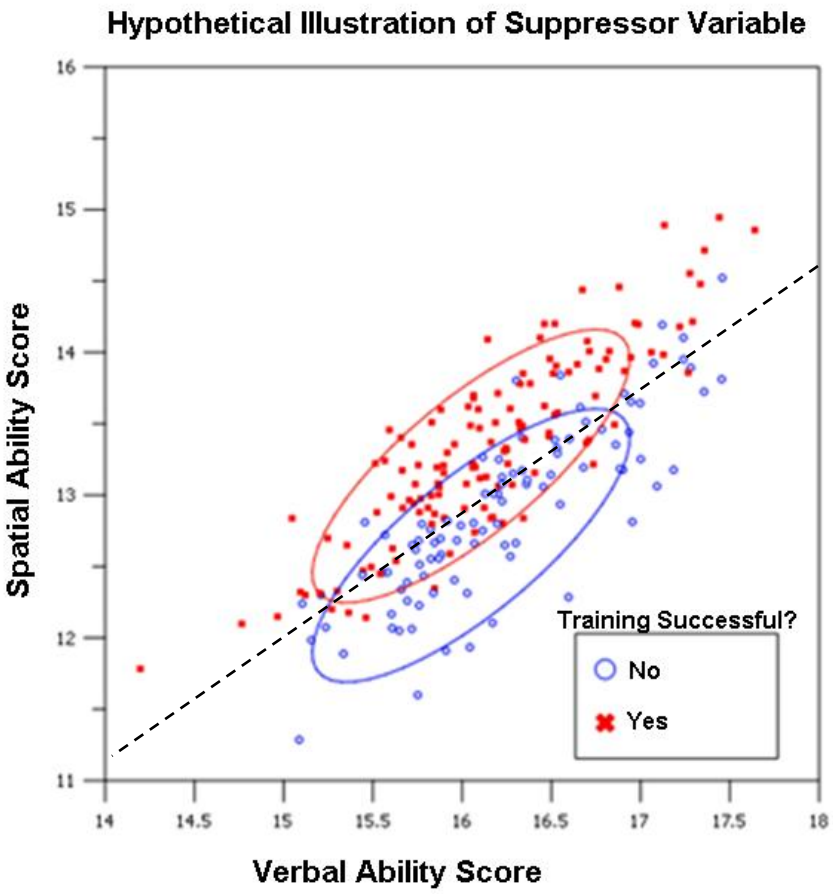


Fig. 1C.

Fig. 2A: X_2 is a valid predictor

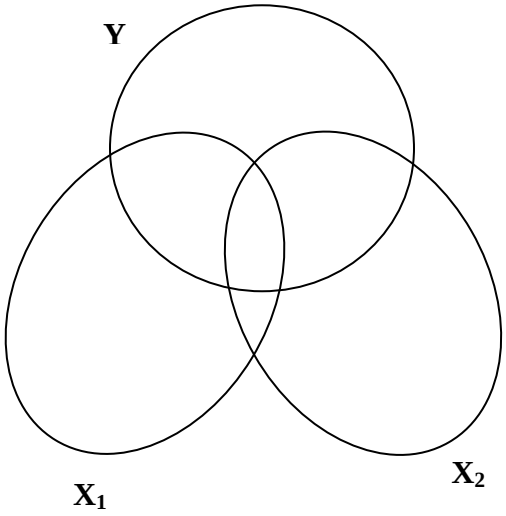


Fig. 2B: X_3 is an irrelevant predictor

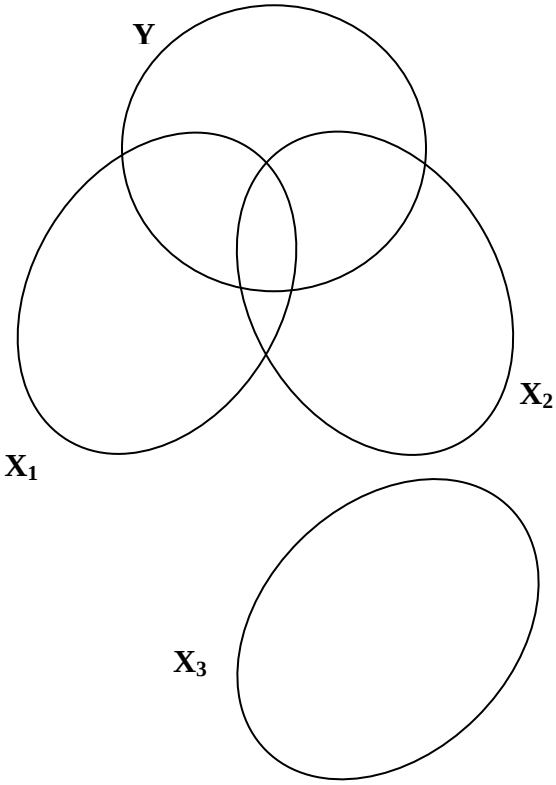


Fig. 2C: X_4 is an extraneous predictor

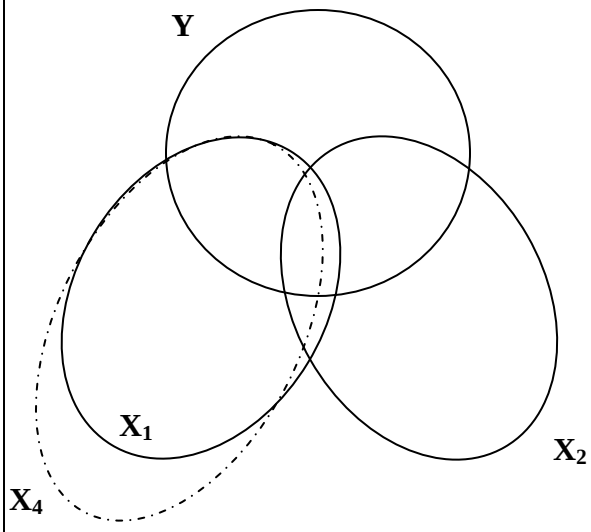
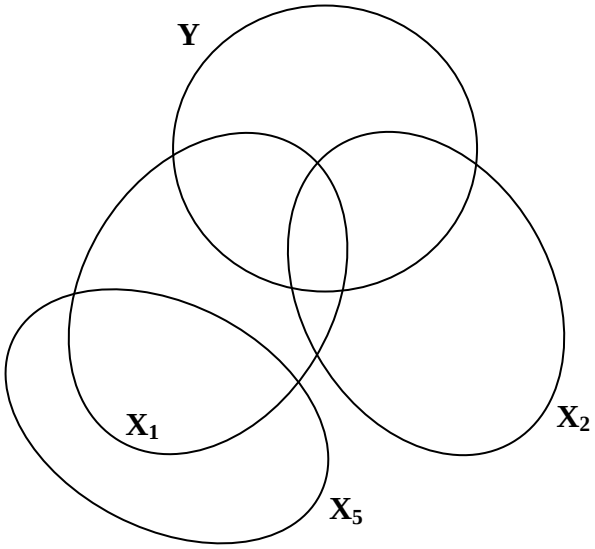


Fig. 2D: X_5 is a (classical) suppressor variable



A1. Suppressor Variables with Gene Expression Data

Suppressor variables, called "proxy genes" in genomics (Magidson, et al., 2010), have no direct effects, but improve prediction by enhancing the effects of genes that do have direct effects "prime genes". Suppressor variables commonly occur with gene expression and other high dimensional data, and often turn out to be among the most important predictors. In addition, because of their sizable correlations with the associated focal predictor, suppressor variables often possess a structural relationship with the focal predictor that can help in interpreting results.

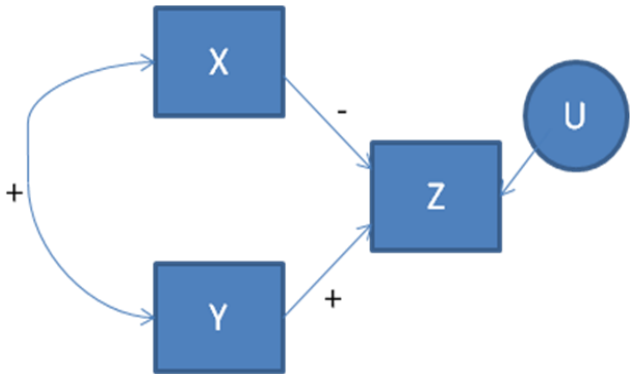


Fig. 3. Path diagram showing the relationship between the prime predictor Y, the suppressor variable X, and the dichotomous dependent variable Z

The path diagram in Fig. 3 indicates:

- Z is dependent variable
- X and Y are predictors of Z
- X (suppressor) has + correlation with Y
- Y has + correlation with Z

Since X is uncorrelated with Z, the partial effect of X on Z controlling for Y is negative.

If it is known that X is a pure suppressor, a more efficient estimate for the partial effect of Y on Z might be obtained using a 2-step approach:

Step 1: Regress Y on X

Step 2: Regress Z on the orthogonal component $y._X$

Hanczar, et al. (2007) showed that 'synergistic gene pairs' are commonplace in gene expression data, and when included in a model, provide improved discrimination between cancer and

normal tissue. Magidson and Wassmann (2010) showed that the reason that Hanczar's gene pairs predict well is that one gene in the pair is a suppressor variable, enhancing the effect of its paired counterpart, and that inclusion of one or more suppressor variables among model predictors improves both predictive performance and reliability across many types of cancer as well as in survival models.

A simple example involving a sample of men with Prostate Cancer (CaP) and an age-matched sample of normals demonstrate the prime/proxy relationship.

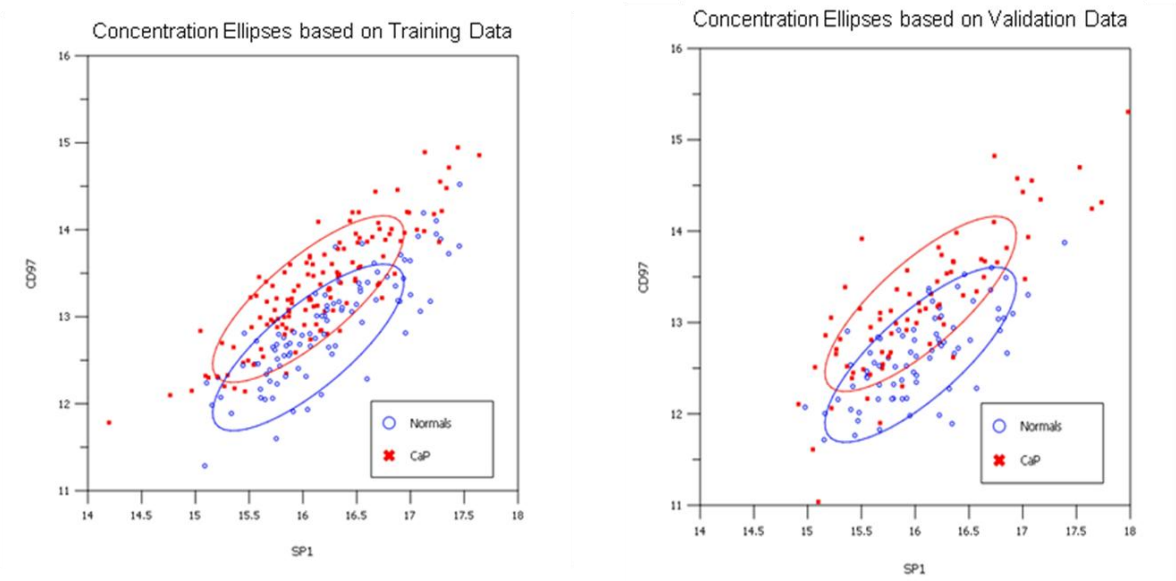


Fig. 4: Example of Prime/Proxy Gene Pair Providing Good Separation of CaP vs. Normals as Confirmed by Validation Data

CaP Subjects have elevated CD97 Δct level as compared to Normals – Red ellipse lies above blue ellipse. CaP and Normals do not differ on SP1, despite its high correlation with CD97.

Inclusion of SP1 significantly improves prediction of CaP vs. Normals over CD97 alone: AUC = .87 vs. .70 (training data), and .84 vs. .73 (validation data).

Figure 4 displays scatterplots of gene expression values for CD97 and SP1 (in delta ct units) from the training and validation data, with separate 68% concentration ellipsoids superimposed for the cancer and normal subjects. From a visual perspective, the validation data supports the hypothesis developed from the training data that the concentration ellipses provide good separation of CaP and Normals. Most normals are contained within the blue ellipse while most CaP subjects are contained within the red ellipse. Although the mean value for SP1 is identical in both groups, classification based on both CD97 and SP1 is improved significantly over prediction based solely on CD97 (see "Session3.ReadingSuppressorVars.pdf" Magidson and

Wassmann 2010). The scatterplots also show evidence of a high positive correlation between CD97 and SP1 in both groups, a necessary condition for a prime/proxy relationship.

Assigned Reading:

"Session3.ReadingSuppressorVars.pdf" (Magidson and Wassmann, 2010)

A2. Suppression Algebra in Logistic Regression

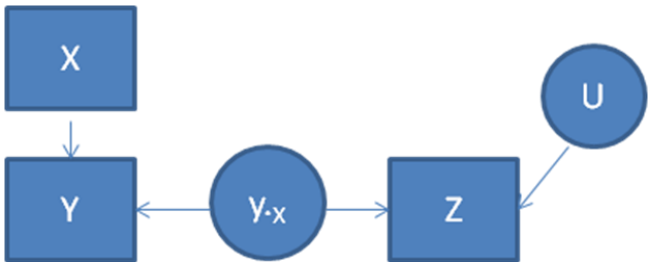


Fig. 5. Alternative path diagram representation of the relationship between the prime predictor Y, the suppressor variable X, and the dichotomous dependent variable Z, where $y._X$ represents the orthogonal component of Y given X.

$X \perp y._X$, $X \perp U$

E(X) = E(Y) = 0

Horst's example: Z is whether or not an individual could be successfully trained as a military pilot, Y is the score on the paper and pencil test measuring the latent construct 'spatial ability', and X is a measure of the independent construct 'verbal ability'. Since the questions on the spatial ability test require both understanding the question (i.e., verbal ability) and visualizing the situations (i.e., spatial ability), the Y is confounded with verbal ability. Thus, persons with a high X score tend to also score higher on Y, despite the fact that X is unrelated to 'spatial ability'. In a regression of Z on Y and X, X serves to remove irrelevant variance from Y, thus boosting its predictive performance (i.e., the orthogonal component $Y._X$ predicts Z better than Y predicts Z)

$$Logit(Z) = \beta y._X + U \qquad Y = bX + y._X \Rightarrow (y._X \equiv Y - bX)$$

$$Logit(Z) = \beta(Y - bX) + U$$

$$\begin{aligned} Logit(Z) &= \beta(Y - bX) + U \\ &= \beta Y - (b\beta)X + U \end{aligned}$$

X is a "pure" suppressor if the coefficient for X in the logistic regression model is the product of b and β, where b = regression coefficient in the linear regression of Y on X and β is the logistic regression coefficient for Y. In this case, the sole role of X in the logistic regression is to remove the irrelevant variance from the predictor Y. Since $y._x$ predicts better than Y, prediction (classification) of Z is improved when the suppressor is included in the model.

For additional information on suppressor variables, see:

Horst, P. (1941). The role of predictor variables which are independent of the criterion. Social Science Research Bulletin, 48, 431-436.

Horst, P. (1966). Psychological measurement and prediction. Belmont, CA: Wadsworth.

Paulhus, et al. (2004). Two Replicable Suppressor Situations in Personality Research. Multivariate Behavioral Research, 39 (2), 303-328. http://neuron4.psych.ubc.ca/~dpaulhus/research/SELF-ENHANCEMENT/downloads/ARTICLES/MBR2004.pdf

B. Simulated Data

For COREXpress Users:

'LMTr&Val.sav'

For XLSTAT-CCR Users:

'LMTr&Val.xls'

Analysis of a Continuous Dependent Variable

Our first simulated data is for a linear regression with a continuous dependent variable Y based on a training sample of size N=5,000. An additional (validation) sample of size N = 5,000 is also available. The file contains of training data ('Validation' = 1) consisting of 100 simulated data sets of size N=50 ('Simulation' = 1-100), and an equal sized validation (test) data set ('Validation' = 1).

The true model consists of 14 'valid' predictors, including the important suppressor variable SP1. In addition to these 14 predictors, available predictors also include 42 *extraneous* predictors (i.e., true coefficients equal zero for these). Fourteen of these extraneous predictors (labeled 'other1-other14') are correlated with the 14 valid predictors, and each of the remaining 28 extraneous predictors (labeled 'extra1-extra28'), is uncorrelated with each of the other 55 predictors.

Theoretically, prediction can never be improved by including any of the irrelevant predictors 'extra1-extra28' in the model, but if some of the valid predictors were excluded, it is possible that prediction can be improved by including one or more extraneous predictors 'other1-other14' that are correlated with the valid predictors excluded.

Table 1A compares regression results obtained from COREXpress (XLSTAT-CCR results will vary slightly) and stepwise regression using the entire training sample. Column 1 lists the subset of predictors for which non-zero coefficient estimates were obtained in at least one of the regressions. The 14 true predictors are listed on top, the first column listing the true population coefficients. The *population* R^2 for these simulated data is approximately .913, and as a benchmark, the R^2 values obtained by applying the true coefficients to the training and independent validation data is .911 and .914 respectively, the slight differences reflecting sampling variation in the two generated samples.

Since only 14 of the 56 predictors are *valid* predictors (i.e., true coefficients are non-zero for these), the true regression is said to be *sparse* (many coefficients equal zero).

Table 1A: Comparison of Results for CCR and Stepwise Regression Models estimated on Training Data (N_{Tr} = 5,000) and Evaluated Using Validation (Test) Data (N_{Val} = 5,000)

		CCR	Stepwise Regression	
	TRUE	K=8	forward	backward
R-sq (Tr) =	0.911	0.911	0.912	0.912
R-sq (CV) =	N/A	0.911		
R-sq (Val) =	0.914	0.913	0.913	0.913
Predictors				
BRCA1	-2.13	-2.2	-2.2	-2.2
CD44	1.85	1.69	1.68	1.68
CD97	1.44	1.45	1.39	1.4
CDKN1A	2.33	2.34	2.34	2.33
EP300	-1.78	-1.64	-1.7	-1.69
GSK3B	4.56	4.59	4.55	4.56
IQGAP1	3.35	3.27	3.33	3.32
MAP2K1	2.75	2.48	2.64	2.73
MYC	-1.81	-1.77	-1.79	-1.77
RB1	-3.82	-3.68	-3.73	-3.75
RP5	5.75	5.8	5.77	5.78
SIAH2	1.15	1.12	1.14	1.14
SP1	-9.55	-9.44	-9.39	-9.39
TNF	2.24	2.25	2.26	2.27
Other1	0	0	0	-0.11
extra4	0	0	0	-0.13
extra5	0	0	0	0.06
extra13	0	0	0	0.05
extra14	0	0	0.06	0.08
extra16	0	0	0	-0.04
extra28	0	0	0	0.06

Column 2 in Table 1A contains results from the K=8-component CCR model (using COREXpress – XLSTAT-CCR results may vary somewhat). As K is reduced in value, the amount of regularization goes up. We selected the model with K=8 components by applying a tuning process based on 1 round of 10-fold cross-validation, results of which are summarized in Table 1B. Table 1A shows that this model, CCR8, correctly yields non-zero coefficients for the 14 valid predictors and correctly excludes all of the extraneous predictors. The remaining columns in Table 1A show that stepwise (backward and forward) regression yields similar

results in terms of the Validation R^2 based on this large sample of N=5,000. However, the stepwise solutions include at least 1 irrelevant predictor in the model.

In contrast to CCR and stepwise regression which are invariant to any linear transformation applied to the predictors, *penalized* regression methods such as lasso require that the predictors be standardized. Lasso also yields results similar to CCR based on this large sample size in terms of validation R^2 but is somewhat worse than both CCR and stepwise regression in terms of predictor recovery, resulting in 23 non-zero coefficients, including 7 irrelevant plus 2 extraneous variables. (The Lasso solution was obtained using GLMNET, tuned using the M-fold cross-validation procedure included in that package).

To determine the number of components K for the CCR model, Table 1B summarizes cross-validation output obtained from COREXpress, for K in the range 2-12. This output includes the cross-validation R^2 statistic (CV-R^2) supplemented by cross-validated predictor counts. Note that CV-R^2 steadily increases as K goes from 2 to 8, and then beginning with K=8 only increases slightly as K increases further for K = 9, 10, 11 and 12. For each K, the bottom row reports the number of predictors that maximize CV-R^2 when that number of predictors is included in the associated K-component model. Note that the correct number P*=14 is reported for K=7-9.

For each predictor, the body of Table 1B (output from COREXpress) reports the number of the 10 CV-Subsamples for which the CCR step-down procedure included that predictor in the model. For example, for CCR8 and CCR9, when the CCR step-down procedure was applied in each of the 10 CV-subsamples (each subsample excluding one of the 10 folds), the 14 true predictors (and only these predictors) were correctly included in the model each and every time, for a total of 10. Table 1B is based on 1 round of 10-folds for 'validation'=0 (N=5,000).

Although the models with K=10, 11, or 12 components yield a higher CV-R^2 than models with K=8 and K=9, CV-R^2 is only slightly higher for the former models, and the predictor counts are less consistent than the latter models, reporting counts less than 10. Thus, by selecting the model with the smallest K among those having (approximately) the highest CV-R^2, we obtain greater consistency in terms of the predictor counts. This type of model selection criterion is similar to that recommended for lasso -- the selected model being the most parsimonious model among those for which the CV error rates are within 1 standard deviation of the lowest CV error rate. (The standard error for the CV-R^2 can be computed and is displayed in the COREXpress (and XLSTAT-CCR) output, when more than 1 round of M-folds is requested.)

Table 1B: Frequency of predictor occurrence in 10 CV-Subsamples for specified K Components

# Components	12	11	10	9	8	7	6	5	4	3	2
CV-R^2	0.9111	0.9111	0.911	0.911	0.9109	0.909	0.8980	0.8911	0.8659	0.81	0.56
BRCA1	10	10	10	10	10	10	10	10	10	10	10
CD44	10	10	10	10	10	10	10	10			
CD97	10	10	10	10	10	10	10	10	10	10	
CDKN1A	10	10	10	10	10	10	10	10	10	10	10
EP300	10	10	10	10	10	9	2				
GSK3B	10	10	10	10	10	10	10	10	10	10	
IQGAP1	10	10	10	10	10	10	10	10	10	10	
MAP2K1	10	10	10	10	10	10	10	10	10	10	
MYC	10	10	10	10	10	10	10	10	10	10	10
RB1	10	10	10	10	10	10	10	10	10		
RP5	10	10	10	10	10	10	10	10	10	10	10
SIAH2	10	10	10	10	10	10	10	10	10	10	10
SP1	10	10	10	10	10	10	10	10	10	10	10
TNF	10	10	10	10	10	10	10	10	10	10	
Other1	10	10	7								
Other10							10	10		10	
Other12		1									
Other13	7	4				1	8	10		10	
Other14		2	2								
extra4	3	9									
extra5		4									
extra14	10	10	1								
# Predictors (P*)	17	18	15	14	14	14	15	15	12	13	6
Total count	170	180	150	140	140	140	150	150	120	130	60

Exercise B1.

For CORExpress Users:

'LMTr&Val.sav'

For XLSTAT-CCR Users:

'LMTr&Val.xls'

In order to provide comparisons to those above but based on 'high-dimensional data' where P>N, in this exercise, we will maintain the P=56 predictors but reduce the sample size by randomly dividing the training sample into 100 equal sized subsamples, each of size N=50.

1) Estimate the 8-component CCR model using 'validation' = 0 and 'simulation' = 1 as the training sample and 10 rounds of 10 folds for cross-validation. What is the CV-R^2? How many predictors are in the selected model? How many of these are valid predictors?

2) Whenever one or more important suppressor variables are included among the predictors, would you expect that the final CCR model would have only K = 1 component? Why or why not?

Exercise B2.

Perform a regression analysis on the same training data using step-wise regression as implemented in SPSS (or XLSTAT). How do the results compare to those obtained from CCR?

Why is it not possible to perform backwards elimination on these data (N=50 and P=56) without imposing an additional condition?

Analysis of a Dichotomous Dependent Variable

For COREXpress Users:

'LDASim.sav'

For XLSTAT-CCR Users:

'LDASim.xls'

Next we will consider an example with a dichotomous dependent variable.

Data were simulated according to the assumptions of Linear Discriminant Analysis. The number of available predictors is $P = G_1 + G_2 + G_3$ where $G_1 = 28$ valid predictors (those with nonzero population coefficients given in Table 2), which include 15 relatively weak predictors (valid predictors with importance scores < .85), $G_2 = 28$ irrelevant predictors (named 'extra1' – 'extra28') uncorrelated with both the dependent variable and with the 28 valid predictors but correlated with each other, and $G_3 = 28$ additional irrelevant predictors ('INDPT1' – 'INDPT28'), each uncorrelated with all other variables. Correlations and variances mimic real data. We generated 100 simulated samples, each consisting of N=50 cases, with group sizes $N_1 = N_2 = 25$.

Table 2: True Linear Discriminant Analysis (LDA) Model Coefficients

Predictors	Unstandardized	Standardized*	Importance	Importance Rank
SP1	-9.55	-5.72	5.72	1
GSK3B	4.56	2.48	2.48	2
RB1	-3.82	-2.30	2.30	3
IQGAP1	3.35	2.13	2.13	4
BRCA1	-2.13	-1.36	1.36	5
TNF	2.24	1.32	1.32	6
CDKN1A	2.33	1.29	1.29	7
MAP2K1	2.75	1.20	1.20	8
MYC	-1.81	-1.19	1.19	9
EP300	-1.78	-1.15	1.15	10
CD44	1.85	1.03	1.03	11
CD97	1.44	0.92	0.92	12
SIAH2	1.15	0.87	0.87	13
MAPK1	1.64	0.79	0.79	14
RP5	1.94	0.76	0.76	15
S100A6	1.22	0.74	0.74	16
ABL1	1.44	0.73	0.73	17
NFKB1	1.22	0.70	0.70	18
MTF1	-1.01	-0.62	0.62	19
CDK2	1.20	0.61	0.61	20
IL18	-0.79	-0.56	0.56	21
PTPRC	-0.98	-0.53	0.53	22
SMAD3	-0.57	-0.35	0.35	23
C1QA	-0.29	-0.30	0.30	24
TP53	0.45	0.26	0.26	25
CDKN2A	-0.31	-0.23	0.23	26
CCNE1	-0.21	-0.19	0.19	27
ST14	-0.18	-0.14	0.14	28

*Standardized coefficient = Unstandardized coefficient multiplied by standard deviation of predictor

Exercise B3.

Follow the steps in Tutorial 1 ("Session3Tutorial1.pdf") for the training sample consisting of the 100 cases in simulation = 1 and simulation = 2. Repeat the analysis, selecting a different training sample of 100 cases. How do the results differ between the 2 samples?

For COREXpress Users:


"Session3Tutorial1.pdf"

For XLSTAT-CCR Users:


"XLCCRtutorial1.pdf"

C. Failure of common prescreening methods to capture suppressor variables

Despite the extensive literature documenting the strong enhancement effects of suppressor variables (e.g., Horst, 1941, Lynn, 2003, Friedman and Wall, 2005), most pre-screening methods omit proxy genes prior to model development, resulting in suboptimal models. This includes several popular approaches such as the prescreening suggested by Bair et al. (2006) in sparse principle components analysis, and the SIS procedure proposed by Fan and Lv (2008) which is used to justify the popular "sure independence screening" results. See page 12 of Session 2 Assigned Reading: "Session 2 Assigned Reading - AMSTAT 2 Paper.pdf" for further discussion.

A useful distinction, provided by Fan and Lv (2008), is between high and ultra-high dimensional data. With ultra-high data, when enough irrelevant variables are included as predictors, the

effects of valid predictors may become overwhelmed. This can occur even in a 1-component CCR model. In order that important suppressor variables show up in component 2, component 1 must contain sufficient correlation with the predictors being suppressed. Therefore, screening may be necessary to eliminate irrelevant variables from component 1. Fan and Lv recognized that their SIS screening approach had the problem of eliminating suppressor variables. Therefore, they proposed ISIS to pre-screen predictors in ultra-high dimensional data where suppressor variables may be present. Fan et al. (2009) present ISIS simulation results based on 3 prime predictors and one suppressor variable which shows a large improvement over SIS.

In Session 4 we will examine the screening option that has been implemented in COREXpress and XLSTAT-CCR. For an example of how this screening approach may be expected to improve over ISIS, see the figure on page 14 of Session 2 Assigned Reading: "Session 2 Assigned Reading - AMSTAT 2 Paper.pdf".

D. Coefficient Path Plots

The coefficient path plot is a good way to visualize the effects of regularization. This plot traces the path of each regression coefficient for increasing (or decreasing) amounts of regularization. For comparability of predictors, coefficients are usually plotted in standardized units (standardized coefficients).

Fig. 6 displays a coefficient path plot for various CCR models which distinguish between men with and without prostate cancer (Magidson and Wassmann, 2010). At one extreme is the Naïve Bayes model (K=1) which imposes an extreme amount of regularization. Since suppressor variables are uncorrelated with the dependent variable, the coefficients for a 1-component CCR model will be 0 for suppressor variables as well as for irrelevant predictors. Note that the coefficient for the suppressor SP1 is close to 0 in Fig. 6 for K=1.

At the other extreme is the traditional saturated logistic regression (or LDA) model. With 10 predictors, this model is equivalent to the 10-component CCR model, the saturated CCR model that contains no regularization at all. As can be seen in Fig. 6, there is little change in any of the coefficients as increasing amounts of regularization are imposed upon the saturated model as K is reduced from 10 to K=9,8,7 and 6. The changes in coefficients take place primarily in the range K = 1-6 components. In particular, substantial change is evident in the coefficient for the suppressor variable SP1. For K>3, the magnitude of the coefficient for SP1 is largest among all predictors.

Irrelevant variables tend to have 0 coefficients for any value of K. Thus, the coefficient path plot is one way to distinguish suppressor variables from irrelevant variables. Another way to identify suppressor variables is to examine the component loadings associated with the first 2 components of a CCR model. The loading on component 1 should be close to 0 while the loading on component 2 should be non-zero.

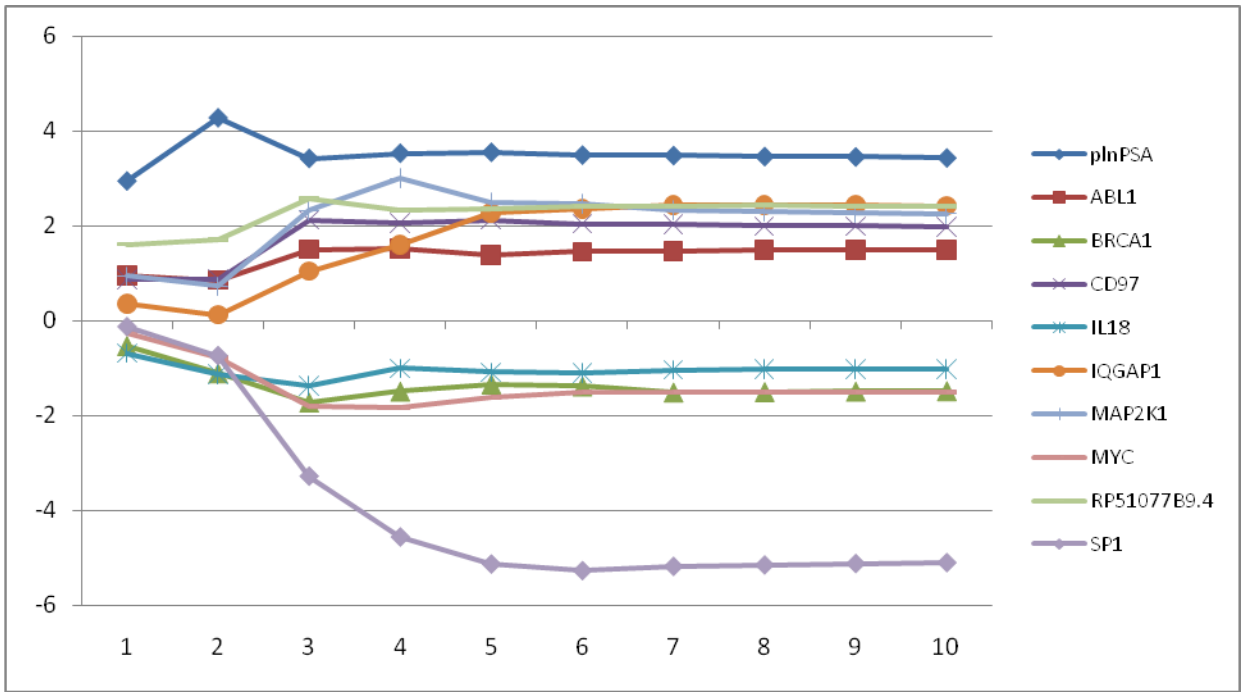


Fig. 6:

Table 3: Comparison of Coefficient Estimates in K-component Model as K goes from 1 to P

# Components =	K = 1	K = 2	K = 3	K = 4	K = 5	K = 6	K = 7	K = 8	K = 9	K = 10
plnPSA	2.95	4.27	3.42	3.53	3.55	3.49	3.48	3.46	3.45	3.43
ABL1	0.95	0.85	1.50	1.51	1.38	1.45	1.46	1.49	1.49	1.49
BRCA1	-0.54	-1.11	-1.72	-1.49	-1.34	-1.38	-1.51	-1.50	-1.49	-1.48
CD97	0.87	0.88	2.12	2.06	2.12	2.04	2.03	1.99	1.99	1.98
IL18	-0.69	-1.13	-1.38	-1.01	-1.09	-1.11	-1.04	-1.03	-1.02	-1.02
IQGAP1	0.35	0.12	1.04	1.59	2.27	2.36	2.43	2.44	2.43	2.42
MAP2K1	0.94	0.72	2.32	3.00	2.50	2.47	2.33	2.29	2.26	2.25
MYC	-0.28	-0.78	-1.80	-1.83	-1.63	-1.51	-1.50	-1.51	-1.51	-1.50
RP51077B9.4	1.61	1.70	2.58	2.33	2.35	2.42	2.42	2.44	2.42	2.41
SP1	-0.14	-0.75	-3.29	-4.57	-5.14	-5.28	-5.19	-5.16	-5.13	-5.11