



# Rethinking Regression, Prediction and Variable Selection in the Presence of High Dimensional Data: Correlated Component Regression

Jay Magidson, Ph.D.

Statistical Innovations Inc.  
[www.statisticalinnovations.com](http://www.statisticalinnovations.com)

Presented on May 25, 2011  
Modern Modeling Methods (M<sup>3</sup>) Conference  
Univ. of Connecticut  
Location: Gentry 144



## **Abstract:**

Recent advances with high-dimensional data show how reliable predictions can be attained from regression models even when the number of predictors exceeds the sample size. We introduce a promising new method called Correlated Component Regression (CCR), and a related step-down algorithm for reducing the number of predictors, provide insights into why CCR works, and compare its performance with stepwise regression based on data generated under traditional linear regression assumptions. The results suggest that a major reason that CCR works so well may be its high power to capture effects of suppressor variables when such variables are among the candidate predictors.



# Outline of Topics

Scope: Regression with a *single dependent variable*  $Y$  and many correlated predictors

- Problems due to high dimensional data
- Approaches for dealing with these problems
- Variable reduction (sparse) methods
- Challenge of retaining suppressor variables as predictors
- Example with simulated data: CCR vs. Stepwise regression
- CCR Extensions
- Future research directions

# Linear Regression Model Assumptions

- Linear regression model expresses the conditional expectation of  $Y$  given  $X$ , denoted ' $E(Y|X)$ ' as a linear function of  $X = (X_1, \dots, X_p)$ .
- Each observation in a sample is generated by an underlying process described by the equation:

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip} + \varepsilon_i \quad i = 1, 2, \dots, N$$

- The model is '*Linear in the parameters*' (i.e.,  $X$ 's may contain non-linear terms – e.g.,  $X_{i3} = X_{i1} * X_{i2}$ )
- $\varepsilon_i$  is random error (random '*disturbance*')
  - $\varepsilon_i$  assumed to be normally distributed
  - Implies that  $Y_i$  is continuous,  $Y|X$  is conditionally normal

# Linear Regression Assumptions

- **Homoscedasticity:** Each disturbance  $\varepsilon_i$  has the same finite variance  $\sigma^2$ .
- **Nonautocorrelation:** Each disturbance  $\varepsilon_i$  is uncorrelated with every other disturbance,  $\varepsilon_j$ .
- **Conditional Expectation of  $\varepsilon_i = 0$ :**  $E[\varepsilon_i \mid x_{j1}, x_{j2}, \dots, x_{jp}] = 0$ .  
Expected value of disturbance at observation  $i$  is not a function of the  $x$ -variables at any observation.

# Least Squares (OLS) Regression

- Ordinary Least Squares regression finds estimates of the  $\beta$ s such that the sum of squared residuals is minimized.
- Least squares is computationally convenient, and has some nice properties.
  - Least squares maximizes  $R^2$
  - Among all unbiased estimators, OLS has minimum variance (BLUE: **B**est **L**inear **U**nbiased **E**stimator).
- Caveats
  - The assumptions must be valid.
  - Some biased estimators may work better (lower prediction error) than unbiased estimators.

# Matrix Formula for OLS Estimator and Implications for its Variance

- $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$
- $\text{Var}[\mathbf{b} | \mathbf{X}] = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$
- For standard error estimate of gth coefficient, replace  $\sigma^2$  by its estimate  $s^2$  and compute the square root of the gth diagonal element of the estimated Var matrix.
- $\text{Var}(b_g) = [1/(N-P)] * (s_Y^2 / s_g^2) * (1-R^2) / (1 - R_g^2)$ , where  $s_Y^2$  and  $s_g^2$  are the variances for Y and  $X_g$ ,  $R^2$  is the regression  $R^2$  and  $(1-R_g^2)$  is 'tolerance' of  $X_g$ ,  $R_g^2$  being for regression of  $X_g$  on the other predictors.
- Lower correlation among Xs yields higher tolerance for some  $X_g$ , and lower standard errors for some  $b_g$ .

# Assumptions – X has Full Rank

- $X$  = predictor matrix of dimension  $N \times P$
- Assume no exact linear relationship exists among any of the  $P$  predictor variables in the model.
- This assumption implies  $N > P$ .
- When there are **exact** co-linearities,  $X'X$  is singular and traditional OLS estimation cannot occur.
- With **near**-collinearity of  $X$  (near singularity of  $X'X$ ), get imprecise, unstable parameter estimates.

# The High-Dimensional Data ( $P > N$ ) Problem in Regression Modeling

## **Problem:**

When the number of predictor variables  $P$  approaches or exceeds sample size  $N$ , coefficients estimated using traditional regression techniques become unstable or cannot be uniquely estimated due to multicollinearity (singularity of the covariance matrix), and in logistic regression, perfect separation of groups occurs in the analysis sample. The apparent good in-sample performance often is due to *overfitting*, and will not generalize to new cases as well as more parsimonious approaches.

The problem is created by the correlation among the predictors that steadily increases as the number of predictors approaches the sample size. With  $P = N$ , any predictor variable can be expressed as an exact linear combination of the other predictors (perfect multicollinearity), even if these predictors are completely uncorrelated in the population. Traditional methods can not handle this problem.



# Near Collinearity Can Occur Even with *Independent* Predictors

Consider high-dimensional data

- $P$  approaching or exceeding  $N$  defines *high-dimensional data*.
- As  $P$  approaches  $N$ , the data becomes increasingly collinear.
- If  $P \geq N - 1$ , *perfect* collinearity occurs. *Regardless of true population correlations, all predictors have 0 tolerance* -- Any predictor  $g$  can be expressed as an exact linear combination of the others:  $R^2_g = 1$ ,  $g = 1, 2, \dots, P$ .

- Suppose predictors are *independent* in the population.
- Consider 5 mutually independent variables:  $X_1, X_2, X_3, X_4, X_5$   
Results\* from regression of  $X_5$  on the other 4 predictors:
- For example, with  $N = 6$ , this *high-dimensional data* ( $P = 4$ ) yields estimated  $R^2_5 = .86$ , even though population  $R^2_5 = 0$ .

Notes: For  $N=9$ , *adjusted*  $R^2 = .452$ , still substantially higher than 0.

- Cross-validated  $R^2$ , CV- $R^2$ , is more stable than *adjusted*  $R^2$
- CV- $R^2$  is consistently less than twice its standard error.

\*Results obtained from demo data set LDASim.sav ;  $INDPT5 = f(INDPT1-INDPT4)$ ; cases selected using 'ID'  $\leq N$ .  
CV results based on 10 rounds of M-folds, with  $M = 5$  for  $N > 10$ , and  $M = 3$  for  $N = 6$  and  $N = 9$ .

N	Estimated $R^2$		Cross-validated (CV)	
	Unadjusted	Adjusted	CV- $R^2$	(std. err.)
5000	0	-0.001	0.0011	(0.0007)
500	0.01	0.008	0.0005	(0.0005)
100	0.05	0.008	0.0006	(0.0008)
50	0.13	0.05	0.0152	(0.0167)
20	0.21	-0.003	0.0268	(0.0385)
10	0.45	0.001	0.0744	(0.0626)
9	0.73	0.452	0.1507	(0.1207)
6	0.86	0.291	-----	-----
5	1	-----	-----	-----



# The Goal is to Reduce Prediction Error

Generally, true model  $f(x)$  is unknown. As shown by Hastie, et. al.\*, prediction error at a given  $x_0$  can be partitioned into 3 distinct parts, where  $\hat{f}$  represents predictions obtained from modeling approach:

$$E[(y - \hat{f}(x_0))^2 \mid x = x_0] = \sigma_\varepsilon^2 + [E\hat{f}(x_0) - f(x_0)]^2 + E[\hat{f}(x_0) - E\hat{f}(x_0)]^2$$

= Irreducible Error + Bias<sup>2</sup> + Variance

This is bias in  $\hat{f}$   
Expectation is over all possible samples

Expected value of variation in  $\hat{f}$   
over repeated samples

\*See Hastie, et. al. 2009, p.223

# Regularization in Regression

- Regularization involves imposing one or more model restrictions, which serve to reduce predictor error variance.
- If restriction(s) imposed are true, no bias is created.
- If restrictions are *not* true, bias is created but variance may still be reduced, often resulting in a net reduction in prediction error. This is referred to as the Bias-variance tradeoff.
- Different kinds of regularization:
  - Traditional – set one or more regression coefficients to zero (reducing P directly): eliminating extraneous predictors (true coefficient = 0) maintains unbiasedness and reduces variance, thus reducing prediction error.
  - Penalized regression – restrict magnitude of regression coefficients, biasing them towards zero, but reducing variance -- Ridge Regression, Lasso.
  - Component/Dimension reduction strategies – set effects of higher dimensions to zero, thus reducing variance.
    - Principal Components Regression (PCR)
    - Partial Least Squares Regression (PLS-R)
    - Correlated Component Regression (CCR)



# Same OLS Predictions are Attainable from P Components as from P Predictors

- Each component  $S_g$  is defined as some linear combination of the predictors
- The components should be linearly independent (matrix A is non-singular)
  - $S_{N \times P} = X_{N \times P} A_{P \times P}$
- OLS predictions based on X ( $P < N-1$ ):

$$\begin{aligned}\hat{Y} &= X\hat{\beta} \\ &= X(X'X)^{-1}X'Y\end{aligned}$$

- Predictions based on S:

$$\begin{aligned}\hat{Y} &= S\hat{\gamma} \\ &= S(S'S)^{-1}S'Y \\ &= XA((XA)'XA)^{-1}(XA)'Y \\ &= XA(A'X'XA)^{-1}A'X'Y \\ &= XAA^{-1}(X'X)^{-1}A^{-1}A'X'Y \\ &= X(X'X)^{-1}X'Y\end{aligned}$$

Thus, using  $K < \min(P, N-1)$  components instead of  $P < N-1$  predictors is a form of regularization – dimension reduction

Equalities 3 and 4 follow from the following standard operations with square matrices :

$$(BC)' = C'B'$$

$$(BC)^{-1} = C^{-1}B^{-1}$$

# Questions

- How should the  $P$  components be formed?
- Which of the  $P$  components should be eliminated?
- What statistical techniques should be used to determine  $P^*$ ?

# M-fold Cross-validation for Model Tuning

- Divide sample into M equal groups (folds), recommended M = 5-10.
- Apply a modeling procedure M times, each time omitting one fold.
  - e.g., procedure such as OLS, or may contain 1 or more tuning parameters\*
- Compute performance criteria (loss function) from cases in omitted fold.
  - e.g., compute average CV-R<sup>2</sup> based on all M omitted folds.
- Choose tuning parameters having best performance (smallest error).
- Alternative to information criteria (e.g., AIC, BIC). Often modified in case of ties or insignificant differences to choose more parsimonious solution.

\*Correlated Component Regression (CCR) utilizes 2 tuning parameters:

K = # components and P = # predictors to include in model.

# M-fold Cross-validation for Model Comparison

- Suppose we obtain predictions from 2 or more regression modeling procedures (e.g., traditional OLS vs. 2-component Correlated Component Regression model – CCR2).
- As alternative to information criteria (e.g., AIC, BIC), can use M-fold CV to choose the one with the highest CV- $R^2$ .
- This approach provides similar results to AIC and BIC for low-dimensional data, and can also be computed with high-dimensional data.

# Can Perform R Rounds of M-fold CV

- Estimate standard error for  $CV-R^2$  based on M rounds of M-fold CV.
- Compute  $CV-R^2$  as average across R separate estimates of  $CV-R^2$ .
- Compute standard error as standard deviation of these R estimates.

# Component (Dimension Reduction) Methods

- Component approaches, like penalized approaches, reduce variance, and hence reduce the magnitude of the coefficients.
- Component (Derived Input) Approaches
  - Principal Component Regression (PCR)
  - PLS Regression (PLS-R)
  - Correlated Component Regression (CCR: CORExpress®)

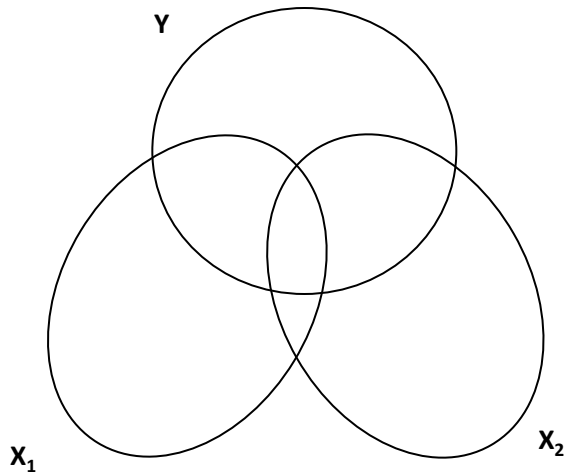
# Sparsity

- In practice, candidate predictor variables may include extraneous or completely irrelevant predictors (population coefficient equals 0). Thus, sparse approaches which exclude certain predictors are favored.
- Component approaches include non-sparse methods such as principal components regression, and the new method correlated component regression (CCR) that includes a sparsity option.

CCR Step-down algorithm can be applied to CCR or to PLS-R

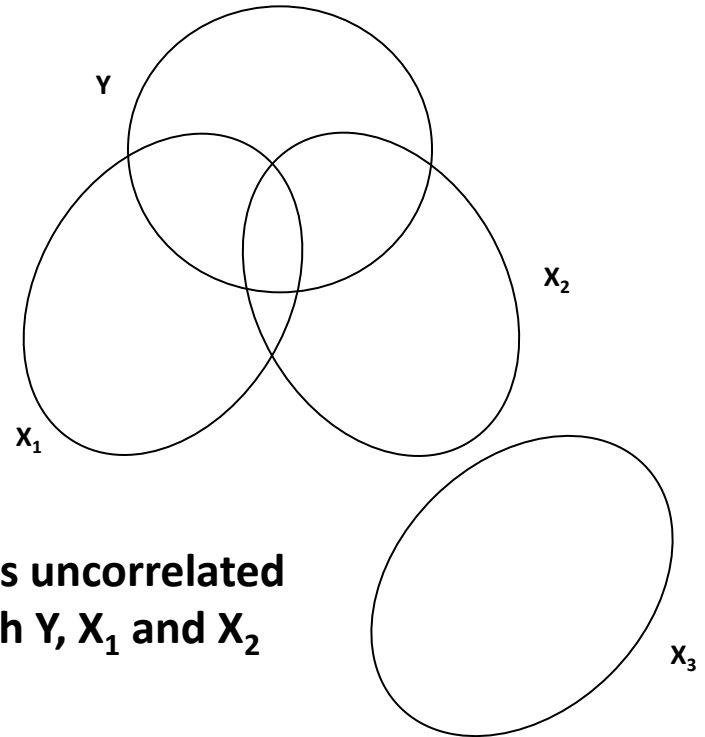
# Valid vs. Irrelevant Predictors

**$X_1$  and  $X_2$  are valid (relevant) predictors**



**Y,  $X_1$  and  $X_2$  are all correlated**

**$X_3$  is an irrelevant predictor**



**$X_3$  is uncorrelated  
with Y,  $X_1$  and  $X_2$**

## Approaches based on Principal Component Analysis (PCA)

1) Principal Component Regression (PCR) – Transform the predictors  $X_1, X_2, \dots, X_p$  to principal components,  $S_1, S_2, \dots, S_p$ , each component defined as a weighted sum of all the predictors. Then use the top  $K < P$  components (those explaining the most predictor variance) as predictors in the model.

- Advantages over stepwise regression:
  - PCR takes into account information on more predictors. There may be only  $K < P$  predictors in model, but each component incorporates information on multiple  $X$  variables, thus *possibly* providing better prediction of  $Y$ .
  - Since components are orthogonal (uncorrelated), problems due to multicollinearity goes away.
- Disadvantages over stepwise regression:
  - The components are not necessarily predictive of  $Y$  and thus may not provide better prediction of  $Y$  than stepwise regression. Example: Component 1 may be completely irrelevant for prediction of  $Y$ .
  - To apply the model, one needs measurements of all  $P$  of the  $X$  variables.

2) Supervised PCR (Bair, et. al., 2006) – Rather than the first  $K$  components, select only the  $K$  components that are significant predictors of  $Y$

- Advantage over PCR – Each component is assured to be predictive of  $Y$ .
- Disadvantage: Excludes components that act as suppressor variables, thus providing poorer prediction than PLS regression and Correlated Component Regression.
- **Non-sparse:** To apply the model, one needs measurements on all  $P$  of the  $X$  variables.

# Principal Components Regression (PCR)

- The most relevant coefficients relate Y to the Xs, not Y to the components.
- Since each component is a weighted sum of the Xs, substituting for the components, one gets coefficients for the Xs.

Example with 2 Components:

$$\hat{Y} = \alpha + b_{1.2}S_1 + b_{2.1}S_2 \quad S_1 = \sum_{g=1}^P \beta_{g.1}X_g \quad S_2 = \sum_{g=1}^P \beta_{g.2}X_g$$
$$\hat{Y} = \alpha + \sum_{g=1}^P (b_{1.2}\beta_{g.1} + b_{2.1}\beta_{g.2})X_g$$



# Partial Least Squares Regression (PLS-R)

- Idea: Replace the  $P$  predictors  $x_g, g=1,2,\dots,P$  by  $K \leq P$  orthonormal\* predictive components  $v_1, v_2, \dots, v_K$

\*orthogonal and standardized to have variance 1 (Y and Xs assumed centered)

- Initialize algorithm: Set  $k=1$  and  $x_g^{(1)} = x_g$  for each  $g$
- Compute  $v_1$ : Each  $x_g^{(k)}$  is weighted by its covariance with  $Y$ , and then divided by the normalizing constant  $s_k$
- Step 1: Compute  $v_k = \sum \text{cov}(y, x_g^{(k)}) x_g^{(k)} / s_k$
- Step 2: For each  $g$ , set  $x_g^{(k+1)} =$  orthogonal component of  $x_g^{(k)}$  with respect to  $v_1, \dots, v_k$  (“deflation” step)
- Step 3: Increment  $k = k+1$  and return to step 1.
- When finished, express each component in terms of original Xs

(“restoration” step): 
$$v_k = \sum_{g=1}^P \lambda_g^{(k)} x_g \quad \hat{y} = \sum_{k=1}^K b_k v_k = \sum_{k=1}^K b_k \sum_{g=1}^P \lambda_g^{(k)} x_g = \sum_{g=1}^P \beta_g x_g$$



# Correlated Component Regression\*

Correlated Component Regression (CCR) utilizes  $K$  correlated components, each a linear combination of the predictors, to predict an outcome variable.

- The first component  $S_1$  captures the effects of predictors which have direct effects on the outcome. It is a weighted average of all 1-predictor effects.
- The second component  $S_2$ , correlated with  $S_1$ , captures the effects of suppressor variables that improve prediction by removing extraneous variation from one or more of the predictors that have direct effects.
- Additional components are included if they improve prediction.

*Prime predictors* (those having direct effects) are identified as those having substantial loadings on  $S_1$ , and *proxy predictors* (suppressor variables) as those having substantial loadings on  $S_2$ , and relatively small loadings on  $S_1$ .

- Simultaneous variable reduction is achieved using a step-down algorithm where at each step the least important predictor is removed, importance defined by the absolute value of the standardized coefficient.  $M$ -fold CV is used to determine the number of components  $K$  and number of predictors  $P$ .



# Example: Correlated Component Regression Estimation Algorithm as Applied to Predictors in Linear Regression: CCR-Im

Step 1: Form 1st component  $S_1$  as average of P 1-predictor models (ignoring  $\alpha_g$ )

$$Y = \alpha_g^{(1)} + \lambda_g^{(1)} X_g + \varepsilon_g^{(1)} \quad g=1,2,\dots,P; \quad \lambda_g^{(1)} = \frac{\text{cov}(Y, X_g)}{\text{var}(X_g)} \quad S_1 = \frac{1}{P} \sum_{g=1}^P \lambda_g X_g$$

1-component model:  $\hat{Y} = \alpha^{(1)} + b_1^{(1)} S_1$

Step 2: Form 2nd component  $S_2$  as average of  $\lambda_g^{(2)} X_g$   
Where each  $\lambda_g^{(2)}$  is estimated from the following 2-predictor model:

$$Y = \alpha^{(2)} + \gamma_{1.g}^{(2)} S_1 + \lambda_g^{(2)} X_g + \varepsilon_g^{(2)} \quad g=1,2,\dots,P; \quad S_2 = \frac{1}{P} \sum_{g=1}^P \lambda_g^{(2)} X_g$$

Step 3: Estimate the 2-component model using  $S_1$  and  $S_2$  as predictors:

$$\hat{Y} = \alpha + b_1^{(2)} S_1 + b_2^{(2)} S_2$$

Continue for  $K = 3, 4, \dots, K^*$ -component model. For example, for  $K=3$ , step 2 becomes:

$$Y = \alpha_g^{(3)} + \gamma_{1.g}^{(3)} S_1 + \gamma_{2.g}^{(3)} S_2 + \lambda_g^{(3)} X_g + \varepsilon_g^{(3)}$$

Final regression coefficients are obtained by OLS regression on components:

$$\hat{Y} = \alpha^{(K)} + \sum_{k=1}^K b_k^{(K)} S_k = \alpha^{(K)} + \sum_{k=1}^K b_k^{(K)} \sum_{g=1}^P \lambda_g^{(k)} x_g = \alpha^{(K)} + \sum_{g=1}^P \beta_g x_g$$

# Some Differences Between PLS-R and CCR ( $K < P$ )

	Invariant to Predictor Scaling?	Components Correlated?
PLS-R	NO	NO
CCR	YES	YES

- As in traditional regression, predictions obtained from CCR are invariant to any linear transformations on the predictors.
- Predictions obtained from PLS-R are not invariant.

# PLS-R is Sensitive to Predictor Scale

Predictions for Y obtained from PLS-R model with  $K < P$  components depend upon the relative scales of the predictors

- If  $x_1$  is replaced by  $x_1^* = cx_1$ , where  $c > 0$ 
  - for  $c > 1$ , 1-component model (PLS1) will tend to have *increased weight* for  $x_1$
  - for  $c < 1$ , 1-component model (PLS1) will tend to have *decreased weight* for  $x_1$

- Example: N=24 car models\*

- Y = PRICE (car price measured in francs)
- $X_1$  = CYLINDER (engine measured in cubic centimeters):
- $X_2$  = POWER (horsepower):
- $X_3$  = SPEED (top speed in kilometers/hour):
- $X_4$  = WEIGHT (kilograms):
- $X_5$  = LENGTH (centimeters):
- $X_6$  = WIDTH (centimeters):

<u>Predictor</u>	<u>Std. Dev</u>
Cylinder	527.9
POWER	38.8
SPEED	25.2
WEIGHT	230.3
LENGTH	41.3
WIDTH	7.7

How do results differ if we use standardized predictors (= Predictor/StdDev)?

\*Data source: Michel Tenenhaus



# For PLS-R, Scale Effects Relative Predictor Importance and Optimal # Components

Implied Relative Importance of Predictors is based on Standardized Coefficients # Components K Determined by Cross-Validated R <sup>2</sup> (CV-R <sup>2</sup> )					
PLS (K=1)		PLS w/ stdzd predictors (K=1)		CCR (K=1)	
Training R <sup>2</sup>	0.74	Training R <sup>2</sup>	0.79	Training R <sup>2</sup>	0.79
CV-R <sup>2</sup>	0.70	CV-R <sup>2</sup>	0.74	CV-R <sup>2</sup>	0.75
Predictors	Standardized Coefficient	Predictors	Standardized Coefficient	Predictors	Standardized Coefficient
CYLINDER	0.73	ZCYLINDER	0.18	CYLINDER	0.18
POWER	0.00	ZPOWER	0.19	POWER	0.19
SPEED	0.00	ZSPEED	0.16	SPEED	0.16
WEIGHT	0.13	ZWEIGHT	0.18	WEIGHT	0.18
LENGTH	0.00	ZLENGTH	0.16	LENGTH	0.16
WIDTH	0.00	ZWIDTH	0.13	WIDTH	0.13
PLS (K=3)		PLS w/ stdzd predictors (K=2)		CCR (K=2)	
Training R <sup>2</sup>	0.83	Training R <sup>2</sup>	0.81	Training R <sup>2</sup>	0.82
CV-R <sup>2</sup>	0.69	CV-R <sup>2</sup>	0.76	CV-R <sup>2</sup>	0.75
Predictors	Standardized Coefficient	Predictors	Standardized Coefficient	Predictors	Standardized Coefficient
CYLINDER	-0.02	ZCYLINDER	0.19	CYLINDER	0.19
POWER	0.43	ZPOWER	0.31	POWER	0.37
SPEED	0.17	ZSPEED	0.22	SPEED	0.20
WEIGHT	0.48	ZWEIGHT	0.18	WEIGHT	0.17
LENGTH	-0.05	ZLENGTH	0.08	LENGTH	0.02
WIDTH	0.00	ZWIDTH	0.01	WIDTH	0.05

Relative importance obtained from PLS-R is sensitive to scaling of predictors (.73 vs. 18).

Additional component required due to scale:  
K\* = 3 (original scale)  
K\* = 2 (standardized)

Overall, importance of CYLINDER goes from unimportant (-.02 with original scale) to important (.19 with standardized).

# CCR Step-down Algorithm

CCR Step-down algorithm can be applied to CCR or to PLS-R

	Non-sparse Version	Sparse Version
PLS-R	Original PLS-R	CCR.pls
CCR	CCR/ no step-down	CCR w/ step-down

When some of the predictors are irrelevant, improved prediction and improved interpretations can be obtained if those predictors are removed from the model.

\*Implemented in CORExpress® program: patent pending regarding this technology

## Correlated Component Regression Step-down Variable Reduction Step

**Step Down:** For a given  $K^*$ , eliminate least *important* predictor in  $K^*$ -component model, where *importance* is quantified by the absolute value of the variable's standardized coefficient, the standardized coefficient computed as the standard deviation times its unstandardized coefficient:

$$\beta_g^* = \sigma_g \beta_g \quad (\text{Can divide by } \sigma_Y \text{ to yield traditional 'beta'})$$

Example with  $K^*=2$ .

Comparing absolute value of standardized coefficients for the  $K^*=2$ -component model determines predictor  $g^*$  to be least important. Then exclude that predictor and repeat the steps of the CCR estimation algorithm on the reduced set of predictors.

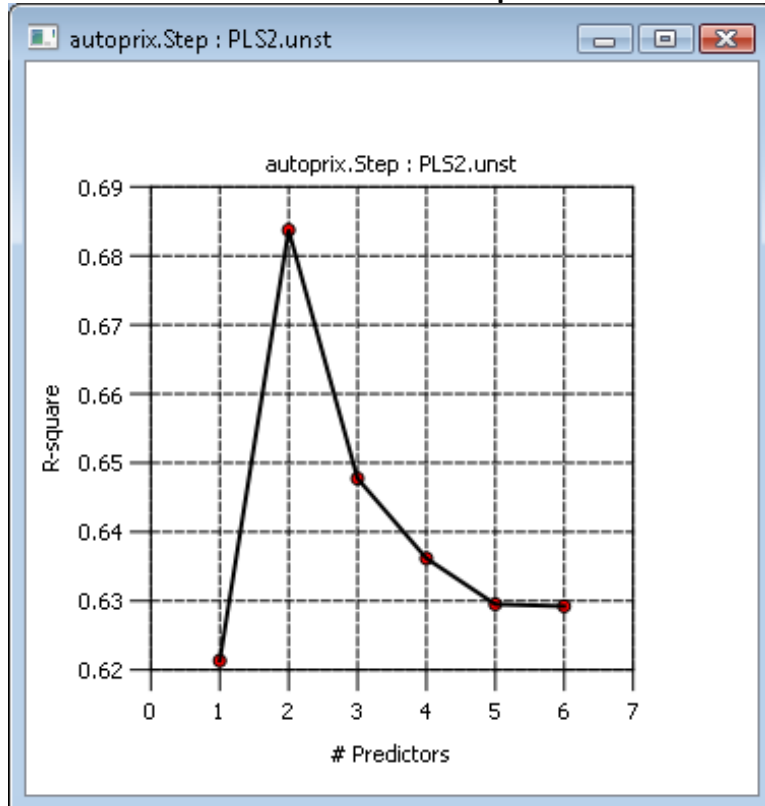
In practice, for large  $P$ , more than 1 predictor can be eliminated at a time. By default, at each step CORExpress eliminates the 1% of the predictors that are least important until  $P < 100$ , at which time it eliminates 1 predictor at a time. This process can continue until 1 predictor remains.

Note: Since  $K$  can never exceed  $P$ :

For  $P = K$ , the model becomes 'saturated' and is equivalent to the traditional regression model. To reduce # predictors further, we maintain saturated model by reducing  $K$  so  $P = K$ . This is similar to traditional stepwise regression with backwards elimination. Thus, for example, for  $K = 4$ , when we step down to 3 predictors, reduce  $K$  so  $K = 3$ . Similarly, when we step down to 1 predictor,  $K=1$ .

# Example: PLS (K=2) with *Unstandardized* Predictors

CV-R<sup>2</sup> as a function of # predictors



Training R<sup>2</sup> 0.74

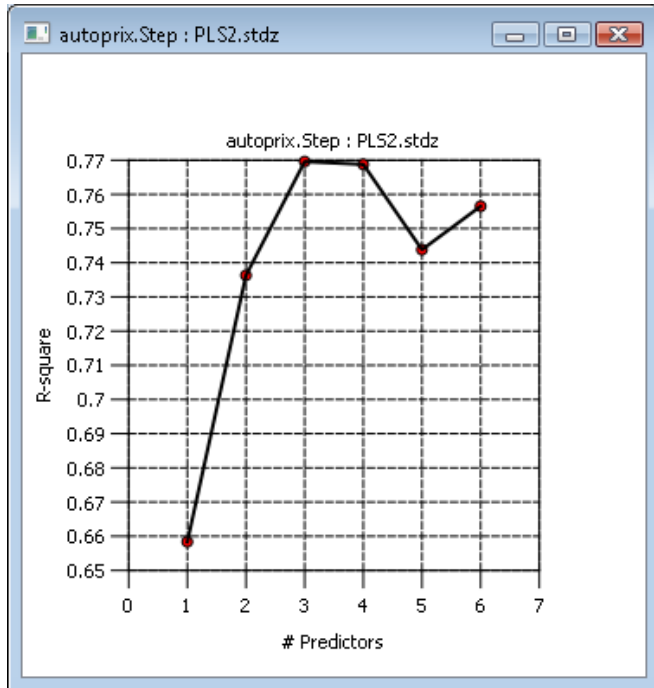
CV-R<sup>2</sup> 0.69 (.05)

Predictors	Standardized Coefficient
CYLINDER	0.64
WEIGHT	0.23

Predictor	All	1	2	3	4	5	6	7	8	9	10
CYLINDER	49	4	6	6	4	6	5	4	5	4	5
WEIGHT	49	4	5	6	5	5	6	4	4	5	5
POWER	28	4	1	6	3	1	1	4	3	3	2
SPEED	6	0	0	6	0	0	0	0	0	0	0
LENGTH	6	0	0	6	0	0	0	0	0	0	0
Total	138	12	12	30	12	12	12	12	12	12	12
Predictors		2	2	5	2	2	2	2	2	2	2

# Example: PLS (K=2) with *Standardized* Predictors

CV-R<sup>2</sup> as a function of # predictors

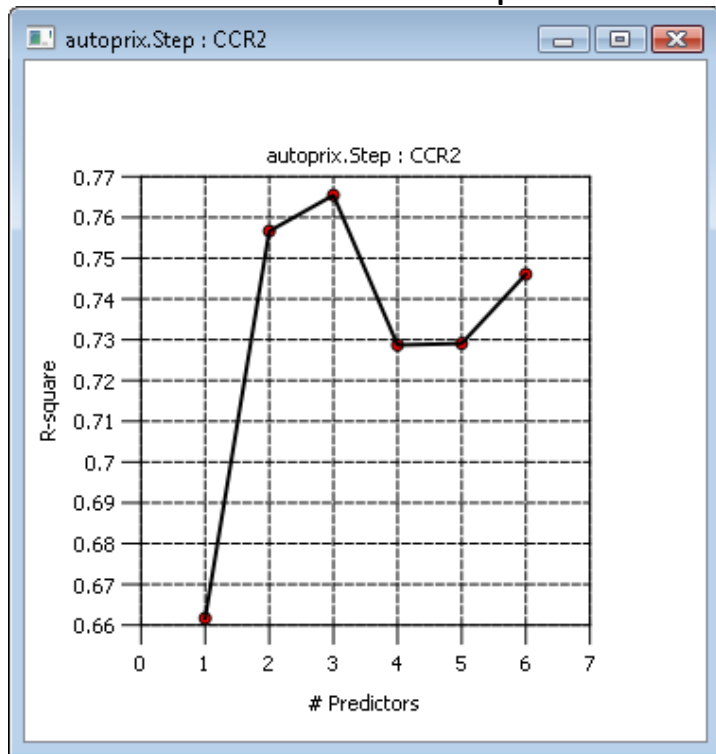


Training R <sup>2</sup>	0.84
CV-R <sup>2</sup>	0.78 (.02)
<b>Standardized Predictors</b>	<b>Standardized Coefficient</b>
ZPOWER	0.58
ZCYLINDER	0.20
ZWEIGHT	0.19

Predictor	All	1	2	3	4	5	6	7	8	9	10
ZPOWER	60	6	6	6	6	6	6	6	6	6	6
ZWEIGHT	60	6	6	6	6	6	6	6	6	6	6
ZCYLINDER	52	5	5	5	5	5	5	6	5	6	5
ZSPEED	25	1	1	5	1	0	1	5	1	5	5
ZLENGTH	7	0	0	2	0	1	0	1	0	1	2
<b>Total</b>	<b>204</b>	<b>18</b>	<b>18</b>	<b>24</b>	<b>18</b>	<b>18</b>	<b>18</b>	<b>24</b>	<b>18</b>	<b>24</b>	<b>24</b>
<b>Predictors</b>		<b>3</b>	<b>3</b>	<b>4</b>	<b>3</b>	<b>3</b>	<b>3</b>	<b>4</b>	<b>3</b>	<b>4</b>	<b>4</b>

# Example: K = 2-component CCR Model

CV-R<sup>2</sup> as a function of # predictors



Training R<sup>2</sup>      **0.84**

CV-R<sup>2</sup>              **0.77 (.03)**

Predictors	Standardized Coefficient
------------	--------------------------

**POWER**            **0.45**

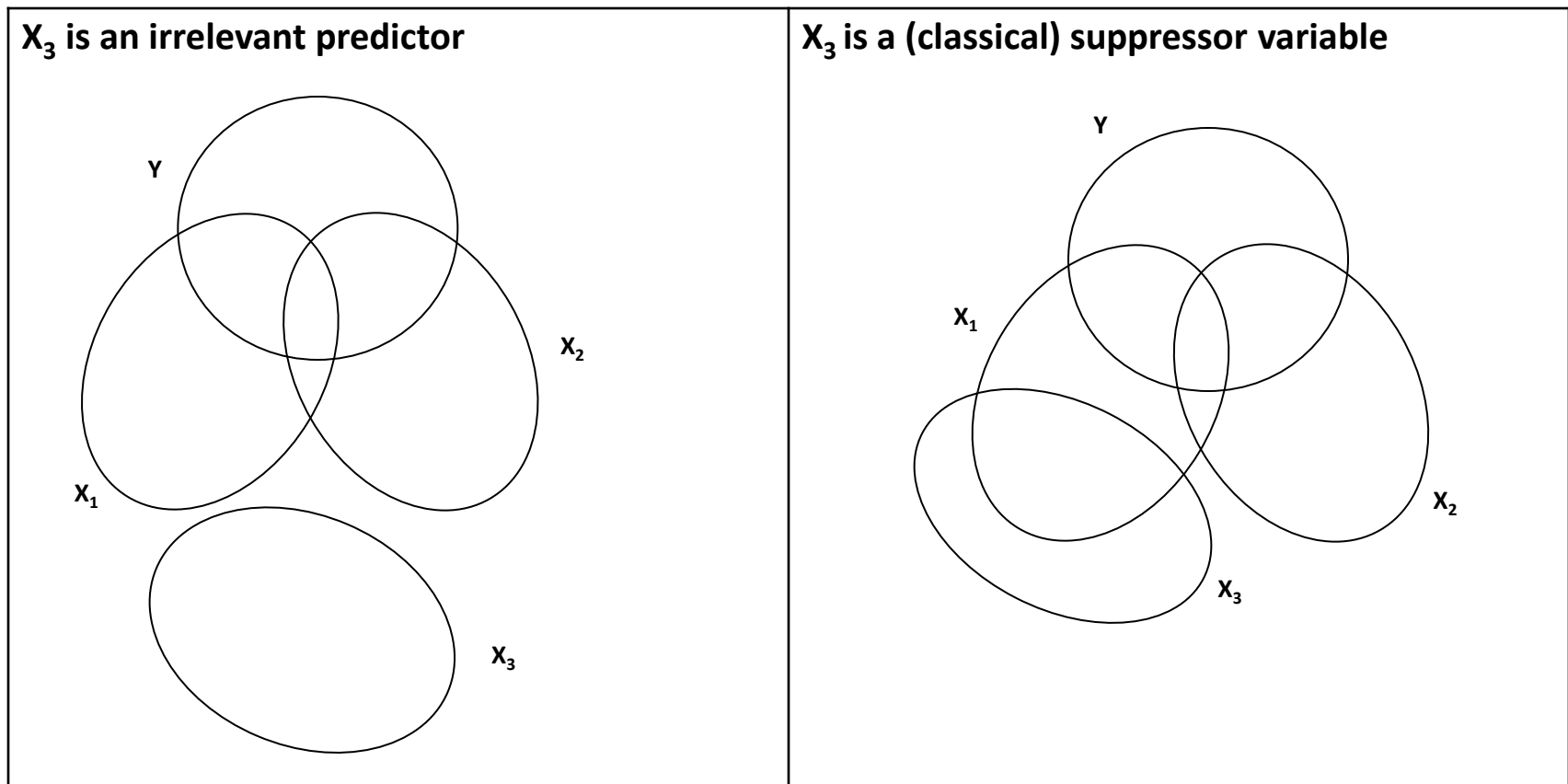
**WEIGHT**          **0.44**

**SPEED**            **0.10**

Predictor	All	1	2	3	4	5	6	7	8	9	10
<b>POWER</b>	<b>60</b>	6	6	6	6	6	6	6	6	6	6
<b>WEIGHT</b>	<b>59</b>	6	6	6	6	6	5	6	6	6	6
<b>SPEED</b>	<b>27</b>	3	6	3	3	2	0	3	4	0	3
<b>CYLINDER</b>	<b>23</b>	2	6	3	2	3	1	3	1	0	2
<b>LENGTH</b>	<b>10</b>	1	6	0	0	1	0	0	1	0	1
<b>WIDTH</b>	<b>7</b>	0	6	0	1	0	0	0	0	0	0
<b>Total</b>	<b>186</b>	18	36	18	18	18	12	18	18	12	18
<b>Predictors</b>		<b>3</b>	<b>6</b>	<b>3</b>	<b>3</b>	<b>3</b>	<b>2</b>	<b>3</b>	<b>3</b>	<b>2</b>	<b>3</b>

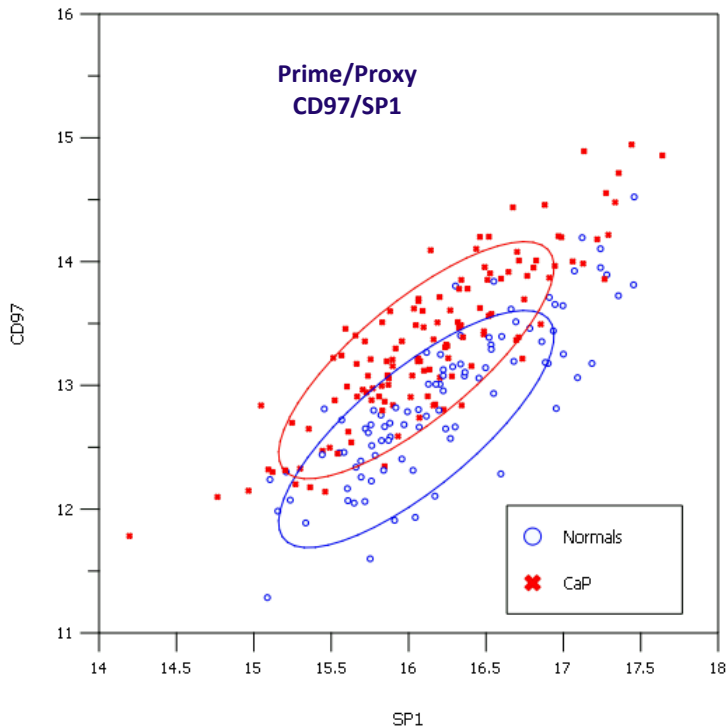
# What is a suppressor variable?

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.



# Example of Suppressor Variable in 2-Gene Model Providing Good Separation of Prostate Cancer (CaP) vs. Normals, Confirmed by Validation Data

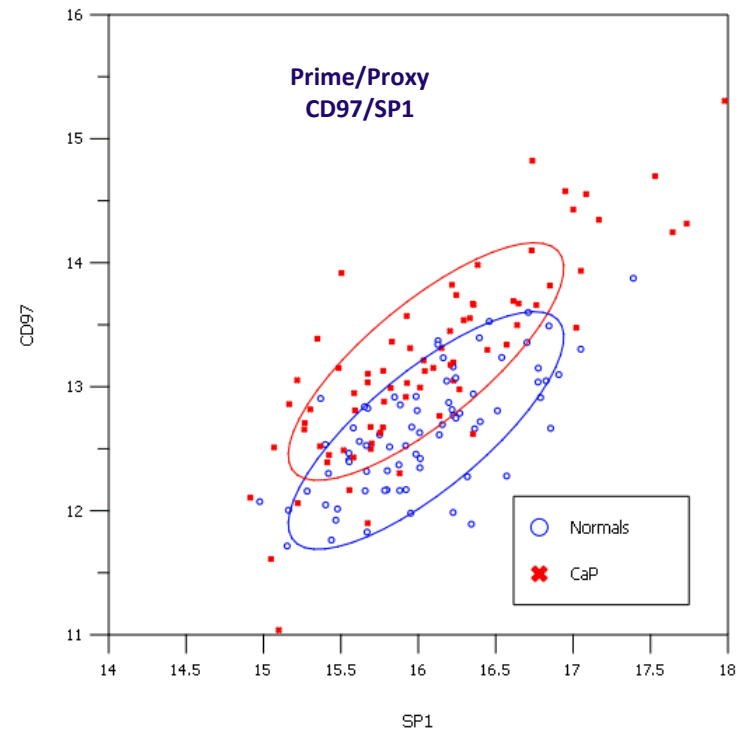
Concentration Ellipses based on Training Data



CaP Subjects have elevated CD97  $\Delta$ ct 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.

Concentration Ellipses based on Validation Data



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) .

See Magidson and Wassmann (2010). “The Role of Proxy Genes in Predictive Models: An Application to Early Detection of Prostate Cancer”, Proceedings from the 2010 Joint Statistical Meetings of the American Statistical Association.

# Important to Retain 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 suppressor variables prior to model development\* resulting in suboptimal models.**

This is akin to: ***“throwing out the baby with the bath water”***.

Suppressors can be identified in a CCR model as follows:

Prime predictors have sizeable loadings on 1st component  $S_1$

Suppressors tend to have zero loadings on component  $S_1$  and sizeable loading on  $S_2$

Note: CCR Components  $S_1$  and  $S_2$  are frequently highly correlated

Since PLS-R components are uncorrelated, typically PLS-R requires more components than CCR to incorporate suppressors in models.

\* For example, supervised principal components analysis (SPCA): Bair, et. al. , 2006;  
SIS: Fan and Lv, 2008.



# Simulation: CCR with Step-down vs. Stepwise Forward Selection

**Design:** Data simulated according to assumptions of Linear Regression\*

- **14 Valid predictors**, including an important suppressor, SP1
- **42 Extraneous predictors** (i.e., true coefficients equal zero for these)
  - ✓ 14 (labeled 'other1-other14') correlated with the 14 valid predictors
  - ✓ 28 (labeled 'extra1-extra28') completely irrelevant
- Continuous dependent variable
- N = 50, 100 simulated samples
- Population  $R^2 \approx .91$

Each method selects  $P^* < 56$  predictors for final model; Each method tuned using M-fold CV. Final models from each method evaluated based on large independent 'test' file (N = 9,750).

Theoretically, prediction can never be improved by including any of the irrelevant predictors 'extra1-extra28' in model, but if some 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 excluded valid predictors.

\* These data and a tutorial are available on website [www.statisticalinnovations.com](http://www.statisticalinnovations.com) and can be analyzed with the demo version of CORExpress®.



# Simulation: CCR with Step-down vs. Stepwise Forward Selection

Large sample results:  $N = 5,000$

Comparison of CCR and Stepwise Regression Models  
 Estimated on Training Data ( $N_{Tr} = 5,000$ ) and Evaluated  
 Using Validation (Test) Data ( $N_{Val} = 5,000$ )

N = 5,000	CCR		Stepwise Regression	
	TRUE	K=8	Forward	Backward
R-sq (Tr) =	0.911	0.911	0.912	0.912
R-sq (Val) =	0.914	0.913	0.913	0.913
	(Unstandardized) Coefficients			
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

K = 8-component CCR model was selected by examining 10-fold CV results for different values for K. This model (CCR8) correctly yields non-zero coefficients for all 14 valid predictors and correctly excludes all of the extraneous predictors.

Stepwise (backward and forward) regression yields similar results in terms of the Validation  $R^2$ . However, the stepwise solutions include at least 1 irrelevant predictor in the model.



# Frequency of Predictor Retention in M = 10 CV-Subsamples for Values of K Ranging from 2-14 Components

# Components	14	13	12	11	10	9	8	7	6	5	4	3	2
CV-R <sup>2</sup> =	0.911	0.911	0.911	0.911	0.911	0.911	0.911	0.909	0.898	0.891	0.866	0.810	0.560
BRCA1	10	10	10	10	10	10	10	10	10	10	10	10	10
CD44	10	10	10	10	10	10	10	10	10	10			
CD97	10	10	10	10	10	10	10	10	10	10	10	10	
CDKN1A	10	10	10	10	10	10	10	10	10	10	10	10	10
EP300	10	10	10	10	10	10	10	9	2				
GSK3B	10	10	10	10	10	10	10	10	10	10	10	10	
IQGAP1	10	10	10	10	10	10	10	10	10	10	10	10	
MAP2K1	10	10	10	10	10	10	10	10	10	10	10	10	
MYC	10	10	10	10	10	10	10	10	10	10	10	10	10
RBL	10	10	10	10	10	10	10	10	10	10	10		
RP5	10	10	10	10	10	10	10	10	10	10	10	10	10
SIAH2	10	10	10	10	10	10	10	10	10	10	10	10	10
SP1	10	10	10	10	10	10	10	10	10	10	10	10	10
TNF	10	10	10	10	10	10	10	10	10	10	10	10	
Other1	10	10	10	10	7								
Other10									10	10		10	
Other12				1									
Other13	7	7	7	4				1	8	10		10	
Other14				2	2								
extra4	3	3	3	9									
extra5				4									
extra14	10	10	10	10	1								
# Predictors =	17	17	17	18	15	14	14	14	15	15	12	13	6
Total count	170	170	170	180	150	140	140	140	150	150	120	130	60

Large sample results: N = 5,000

For these data, the true # components, K = 14, corresponds to the 14 valid predictors. However, better recovery of the true structure occurs with K = 8 or 9.

CV-R<sup>2</sup> increases steadily as K goes from 2 to 8, and then increases only slightly for K = 9-14.

For each K, the bottom row reports the number of predictors that maximize CV-R<sup>2</sup> when that number of predictors is included in the K-component model.

Note that the correct number P\*=14 is reported only for K=7-9.



# Simulation: CCR with Step-down vs Stepwise Forward Selection

Comparison of CCR vs. Stepwise Forward Regression Models Estimated on Simulation #1 (N=50) and Evaluated Using Validation (Test) Data ( $N_{val} = 9,950$ )

N = 50	TRUE	CCR8	Stepwise Forward*	
R-sq (Tr) =	0.97	0.89	0.95	
R-sq (Val) =	0.91	0.71	0.68 Reported	
	Coefficients		p-val	
BRCA1	-2.13	-2.23	-1.51	0.004
CD44	1.85	0	0	
CD97	1.44	2.77	2.92	0.00005
CDKN1A	2.33	3.33	2.15	1.60E-06
EP300	-1.78	-1.60	0	
GSK3B	4.56	0	0	
IQGAP1	3.35	3.57	6.16	5.30E-07
MAP2K1	2.75	0	0	
MYC	-1.81	0	0	
RB1	-3.82	0	0	
RP5	5.75	6.25	6.63	4.40E-12
SIAH2	1.15	0	0.98	0.00023
SP1	-9.55	-8.66	-9.75	1.20E-14
TNF	2.24	2.78	2.43	2.00E-06
Other2	0	0	-1.68	0.009
Other3	0	0	0.56	0.001
Other4	0	0	-2.69	0.024
extra9	0	0	1.12	0.005

## Results from simulation #1 (N=50):

CCR outperforms stepwise regression

- Higher Validation  $R^2$  for CCR (.71 vs. .68)
- Smaller  $R^2$  drop-off from the training data indicating greater reliability:
  - ✓  $.89 - .71 = .18$  for CCR
  - ✓  $.95 - .68 = .27$  for stepwise
- Retains fewer extraneous predictors:
  - ✓ 8 valid and 0 extraneous predictors for CCR
  - ✓ 8 valid plus 4 extraneous for stepwise

Also, p-values (right-most column) reported in stepwise regression output are substantially less than .05 for all predictors, mistakenly suggesting statistical significance. These p-values have a downward bias due to the effects of selection.

Note: Final CCR model is saturated:  $K=P=8$ , yielding predictions equivalent to OLS regression with these 8 predictors.

\*Results from the backward elimination option are not reported because this option cannot be performed with  $P > N$  due to singularity of the covariance matrix.



# Simulation: Summary of Results Across All 100 Simulations for N=50

Overall, across all 100 subsamples, CCR outperformed stepwise regression.

- On average, the CCR model includes:
  - ✓ 2 more valid predictors than stepwise regression (9.0 vs. 7.1)
  - ✓ approximately the same number of extraneous predictors (2.5 vs. 2.2).
- Average correlation with score based on true model and predicted score:
  - ✓ .942 for CCR
  - ✓ .907 for stepwise regression
- Average Mean Squared Error (AMSE)
  - ✓ 1.1 (.079) for CCR
  - ✓ 3.2 (.338) for stepwise
- Final model retained suppressor variable SP1 (most important predictor)
  - ✓ 100% of samples for CCR
  - ✓ 74 % for stepwise regression
- Surprise: In these 74%, AMSE was comparable suggesting that CCR improvement may be largely due to increased power of including suppressors.

# CCR Variants in CORExpress®

CCR-Logistic: Logistic Regression Models

CCR-LDA: Linear Discriminant Analysis

CCR-Cox: Survival (Event History) Models:

Latent Class (Clustering) Applications:

- Selection of predictors prior to LC modeling
- Separate models for each LC segment  
(under development)



# CCR Algorithm for Logistic Regression: CCR-Logistic

Step 1: Form 1st component  $S_1$  as average of P 1-predictor models (ignoring  $\alpha_g$ )

$$\text{Logit}(Z) = \alpha_g + \beta_g X_g \quad g=1,2,\dots,P; \quad S_1 = \frac{1}{P} \sum_{g=1}^P \beta_g X_g$$

1-component model:  $\text{Logit}(Z) = \alpha + \gamma S_1$

Step 2: Form 2nd component  $S_2$  as average of  $\beta_{g.1} X_g$

Where each  $\beta_{g.1}$  is estimated from the following 2-predictor logit model:

$$\text{Logit}(Z) = \alpha_{.1} + \gamma_g S_1 + \beta_{g.1} X_g \quad g=1,2,\dots,P; \quad S_2 = \frac{1}{P} \sum_{g=1}^P \beta_{g.1} X_g$$

Step 3: Estimate the 2-component model using  $S_1$  and  $S_2$  as predictors:

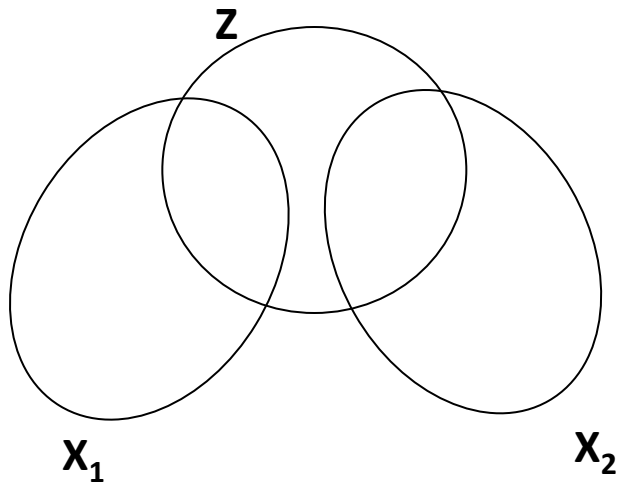
$$\text{Logit}(Z) = \alpha + b_{1.2} S_1 + b_{2.1} S_2$$

Continue for  $K = 3, 4, \dots, K^*$ -component model. For example, for  $K=3$ , step 2 becomes:

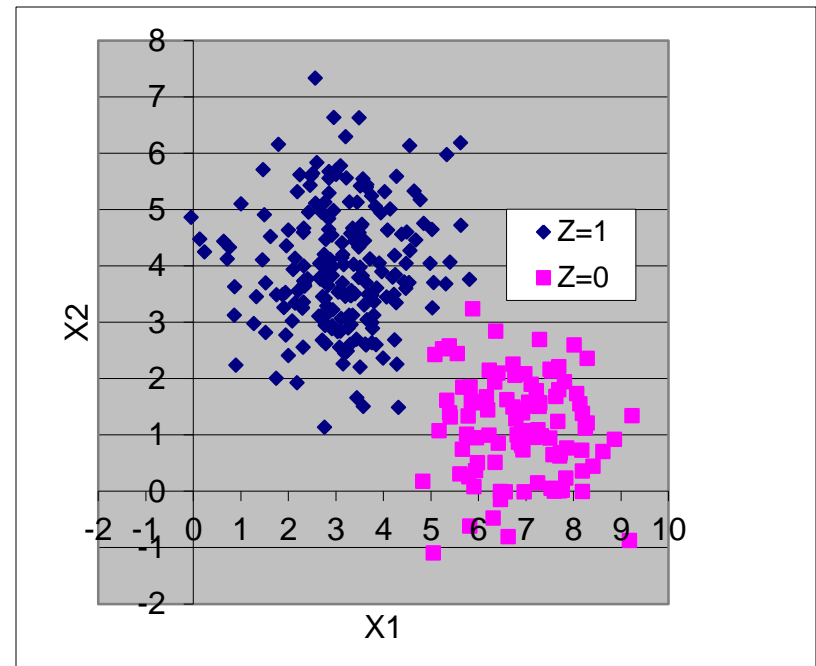
$$\text{Logit}(Z) = \alpha_{.12} + \gamma_{g.1} S_1 + \gamma_{g.2} S_2 + \beta_{g.12} X_g$$

# Equivalence between Naïve Bayes and 1-component CCR-Logistic or CCR-LDA

$X_1$  and  $X_2$  are conditionally independent given  $Z$   
( $X_1 \perp X_2 | Z=1$ ) & ( $X_1 \perp X_2 | Z=0$ )



$X_1$  and  $X_2$  are conditionally independent  
but overall,  $X_1$  and  $X_2$  need not be independent



Naïve Bayes model is very popular in data mining

# Naïve Bayes Works Well with High-Dimensional Data

- With high dimensional data (small samples and many predictors), when data are generated according to the assumptions of linear with discriminant analysis, maximum likelihood estimation based on LDA does not work well. In particular, the simple Naïve Bayes Rule:

*“greatly outperforms the Fisher linear discriminant rule (LDA) under broad conditions when the number of variables grows faster than the number of observations”,* Bickel and Levina (2004)

- Naïve Bayes rule is equivalent to the 1-component CCR model (CCR1).
- Traditional regression is equivalent to a saturated CCR model – CCR with at most  $K=\min(P,N-1)$  components.
- Typically, CCR with 2-8 components (CCR2-CCR8) works best in practice.

Note: Naïve Bayes fails to capture the effects of suppressor variables since by definition, suppressor variables will have zero loadings on component #1.

# Comparisons to Other Sparse Regression Methods

*Sparse* means method involves variable reduction

- A) Sparse Penalty Approaches – dimensionality reduced by setting some coefficients to 0
- LARS/Lasso (L1- regularization): GLMNET (R package)
  - Elastic Net (Average of L1 and L2 regularization): GLMNET (R package)
  - Non-convex penalty: e.g., TLP (Shen, et. al, 2010); SCAD, MCP -- NCVREG (R package)
- B) PLS Regression – dimensionality reduced by excluding higher order components  
P predictors replaced by  $K < P$  *orthogonal components* each defined as a linear combination of the P predictors; orthogonality requirement yields extra components
- e.g., Sparse Generalized Partial Least Squares (SGPLS): SPLS R package  
-- Chun and Keles (2009)



# Results from Full CCR-LDA Simulation -- 100 simulated samples

**Design:** Data simulated according to assumptions of **Linear Discriminant Analysis (LDA)**

$G_1 = 28$  predictors (including 15 weak predictors) plus  $G_2 = 28$  irrelevant predictors

2 Groups:  $N_1 = N_2 = 25$ ; **100 simulated samples**

Method M select  $G^*(M) < 56$  predictors for final model; Each method tuned using validation data with  $N_1 = N_2 = 25$ . Final models from each method evaluated based on large independent 'test' file with  $N_1 = N_2 = 2,500$ .

## **Sparse Regression Methods:**

Correlated Component Regression (CCR), Elastic Net (L1 + L2 regularization, Zou and Hastie, 2005), Lasso (L1 regularization), and sparse PLS regression (sgpls, Chun and Keles, 2009)

### **misclassification error rate:**

CCR (17.4%), sparse PLS (19.3%), Elastic Net (21.1%), lasso (21.6%)

### **Number (%) irrelevant variables:**

CCR (3.4, 23%), lasso (4.3, 31%), Elastic Net (6.6, 34%), sparse PLS (6.9, 34%)

### **% of simulated samples where important suppressor variable included in model:**

CCR (91%), sparse PLS (78%), Elastic Net (61%), lasso (51%)

### **Average # predictors in model:**

lasso (13.6), CCR (14.5), Elastic Net (19.2), sparse PLS (20.4)



# Results from Full CCR-LM Simulation -- 100 simulated samples

**Design:** Data simulated according to assumptions of **Linear Regression**

$G_1 = 14$  valid + 14 extraneous + 28 irrelevant predictors;

Continuous dependent variable,  $N = 50$ , population  $R^2 = .91$ ; 100 simulated samples

Method M select  $G^*(M) < 56$  predictors for final model; Each method tuned using  $N=50$  validation file.  
Final models from each method evaluated based on large independent 'test' file ( $N = 5,000$ ).

TLP = nonconvex (truncated L1) penalty (Shen, et. al., 2010)

**Number of 'True' Predictors included, Percentage of included that were 'True':**

CCR (9.7, 78%), TLP (10.3, 50%), sparse PLS-R (9.5, 48%), Elastic Net (12, 35%)

**# irrelevant *uncorrelated* variables included in model :**

CCR (1.0, 8%), TLP (6.4, 31%), sparse PLS-R (6.4, 33%), Elastic Net (14.1, 41%)

**# irrelevant *correlated* variables included in model:**

CCR (1.8, 15%), sparse PLS-R (4.4, 22%), Elastic Net (8.0, 23%), TLP (4.0, 27%)

**Mean squared error:**

CCR (3.13), sparse PLS-R (3.34), Elastic Net (3.50), TLP (3.55)

# tuning parameters: CCR (3x50), sparse PLS-R (3x50), TLP (5x100), Elastic Net (10x50)

# Future Research: Further Challenge Posed by Ultra-High Dimensional Data

## **Problem and solution:**

For ultra-high dimensional data with many irrelevant predictors, typical with gene expression data, by chance component #1 will contain large loadings for some irrelevant predictors, and small (non-zero) loadings for many other irrelevant predictors, serving to dilute the ability of component #1 to capture the important prime predictors. The ability of component #2 to capture the effects of suppressor variables is dependent upon component #1 measuring the associated prime predictors. To improve the correlation of component #1 with the prime predictors, an initial variable selection 'screening' step may be performed.

Screen option A: Include 2 additional tuning parameters (currently implemented)

- # non-zero loadings on component 1
- # non-zero loadings on component 2



# References

Bair, E., T. Hastie, P. Debnath, and R. Tibshirani (2006). Prediction by supervised principal components. *Journal of the American Statistical Association* 101, 119–137.

Bickel and Levina (2004). Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations, *Bernoulli* 10(6), 989-1010.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96, 1348-1360.

Fan, J. and J. Lv (2008). Sure Independence Screening for Ultra-High Dimensional Feature Space (with Addendum), *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Volume 70, Issue 5, pages 849–911, November.

Fort, G. and Lambert-Lacroix, S. (2003). Classification Using Partial Least Squares with Penalized Logistic Regression. IAP-Statistics, TR0331.

Friedman, L. and M. Wall (2005). Graphical Views Of Suppression and Multicollinearity In Multiple Linear Regression. *American Statistician*, May 2005. Vol 59, No. 2, pp 127-136.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1), 1-22.

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

Hyonho, C. and S. Keleş (2009). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. University of Wisconsin, Madison, USA.



# References (continued)

Lynn, H. (2003). **Suppression and Confounding in Action**. *The American Statistician*, Vol.57, 2003.

Magidson, J. (2010). **User's Guide for CORExpress**. Belmont MA: Statistical Innovations Inc.

Magidson, J. (2010). **A Fast Parsimonious Maximum Likelihood Approach for Prediction Outcome Variables from a Large Number of Predictors**. *COMPSTAT 2010 Proceedings*. Forthcoming.

Magidson, J., and K. Wassmann, (2010) **"The Role of Proxy Genes in Predictive Models: An Application to Early Detection of Prostate Cancer"**, *Proceedings of the American Statistical Association*.

Magidson, J. and Y. Yuan (2010) **"Comparison of Results of Various Methods for Sparse Regression and Variable Pre-Screening"**, unpublished report #CCR2010.1, Belmont MA: Statistical Innovations.

Shen, X., Pan, W., Zhu, Y., and Zhou, H. (2010). **"On L0 regularization in high-dimensional regression"**, to appear.

Vermunt, J.K. (2009): **Event history analysis**. in R. Millsap (ed.) *Handbook of Quantitative Methods in Psychology*, 658-674. London: Sage.

Zou, H. and Hastie, T. (2005). **Regularization and variable selection via the elastic net**. *J. Roy. Statist. Soc. Ser. B* 67, 301-320.