

# Correlated Component Regression A Sparse Alternative to PLS Regression

Jay Magidson  
Statistical Innovations Inc.  
375 Concord Ave., Suite 007  
Belmont, MA 02478 USA  
statisticalinnovations.com

Presented at:  
5th ESSEC-SUPELEC Statistical Workshop on “PLS (Partial Least Squares) Developments”  
Paris, France  
May 20, 2011

# Outline of Topics

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

- Comparison of two methods: CCR (new) and PLS-R
- New sparse algorithm to reduce number of predictors which can be applied to either CCR or PLS-R
- Importance of algorithm retaining suppressor variable predictors
- Example with simulated data: CCR vs. Stepwise regression
- CCR Extensions
- Future research directions

# 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:

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

# Equivalences Between PLS-R, CCR, and OLS

When  $K=P$  (i.e., # components equals # predictors), we have a *saturated* model.  
 OLS-predictions based on *any*  $P$ -components (saturated model) are same as those based on  $X$ .

Proof:

Assume that the components  $S_1, S_2, \dots, S_P$  are linearly independent, so  $A$  has full rank  $P$ .

$$S = X_{N \times P} A_{P \times P}$$

Any new data points  $X_{new}$  can be expressed in terms of the components as  $S_{new} = X_{new} A$ .

The OLS predictions for  $Y$  based on  $X$  are seen to be identical to those based on  $S$ :

$$\text{OLS predictions based on } X: X_{new} (X^T X)^{-1} X^T Y$$

$$\begin{aligned} \text{OLS Predictions based on } S: S_{new} (S^T S)^{-1} S^T Y \\ &= X_{new} A (A^T X^T X A)^{-1} A^T X^T Y \\ &= X_{new} A A^{-1} (X^T X)^{-1} (A^T)^{-1} A^T X^T Y \\ &= X_{new} (X^T X)^{-1} X^T Y \end{aligned}$$

# 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):

<b>Predictor</b>	<b>Std. Dev</b>
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> )					
PLS1 (K=1)		PLS1 w/ stdzd predictors (K=1)		CCR1 (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
PLS3 (K=3)		PLS2 w/ stdzd predictors (K=2)		CCR2 (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).

# Relationships for 1-Component Models

## Unstandardized predictors:

- With  $P = 1$  predictor, model is saturated ( $K=P$ ) so  $CCR1 = PLS1 = OLS$ 
  - ✓ Regression coefficient estimate =  $COV(Y,X)/VAR(X)$
- With  $P > 1$  predictors,  $CCR1$  and  $PLS1$  can differ considerably
  - ✓ Coefficient estimates for  $CCR1$  are proportional to  $COV(Y,X_g)/VAR(X_g)$
  - ✓ Coefficient estimates for  $PLS1$  are proportional to  $COV(Y,X_g)$ , so predictors with larger variance have a larger weight and may dominate

## Standardized predictors:

- ✓ Since  $VAR(X_g) = 1$  for all  $g=1,2,\dots,P$  :  
 $COV(Y,X_g)/VAR(X_g) = COV(Y,X_g)$  and  $CCR1 = PLS1$  (K = 1)

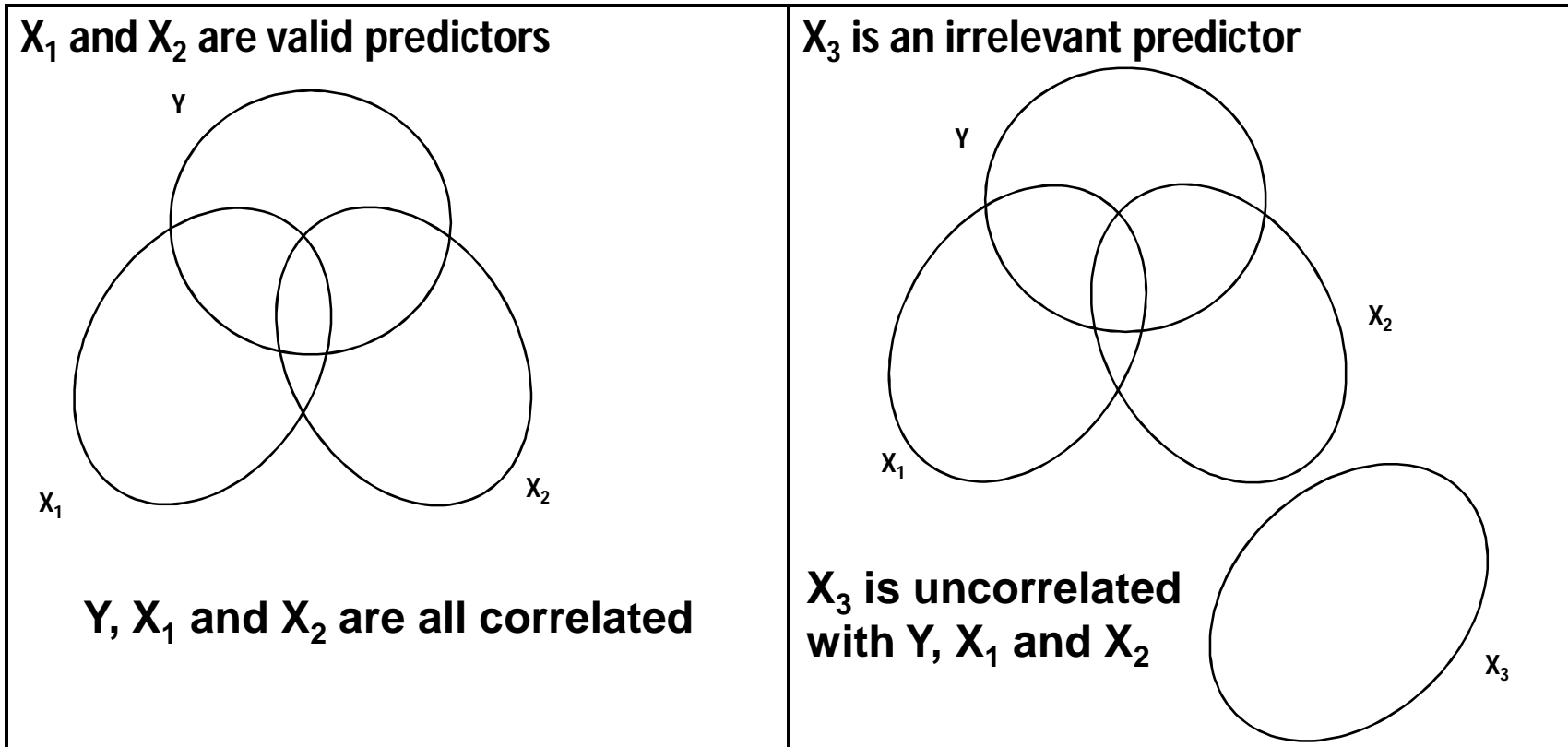
# CCR Step-down Algorithm

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.

# Valid vs. Irrelevant Predictors



## 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:

$$\beta_g^* = \sigma_g \beta_g$$

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,  $K=1$ .

# 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.
  - Procedure may contain 1 or more tuning parameters\*
- Compute performance criteria (loss function) from cases in omitted fold.
  - e.g., for CCR-lm 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:

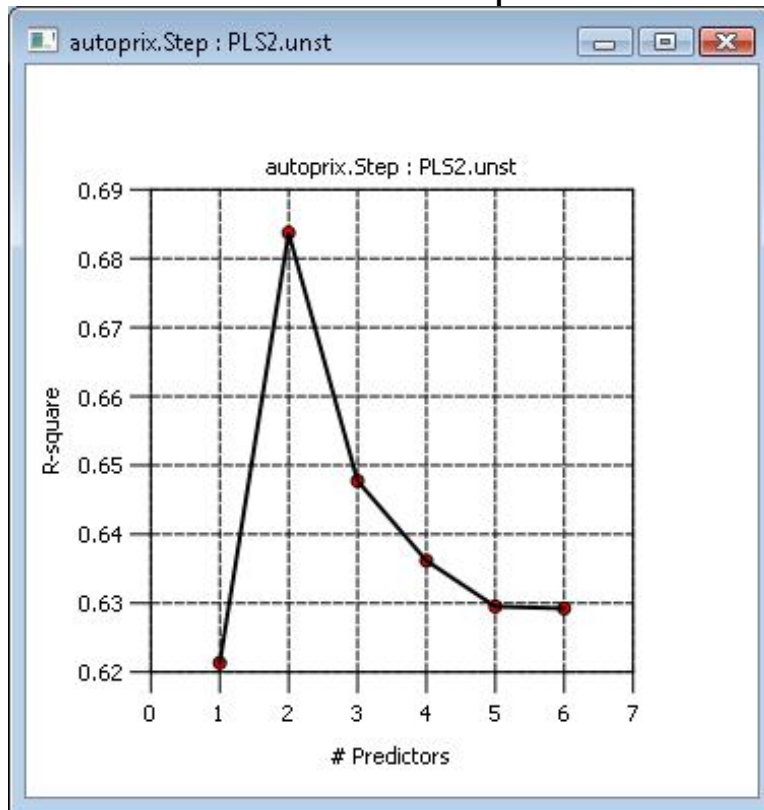
K = # components and P = # predictors to include in model.

# 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.

# Example: PLS2 with Unstandardized Predictors

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

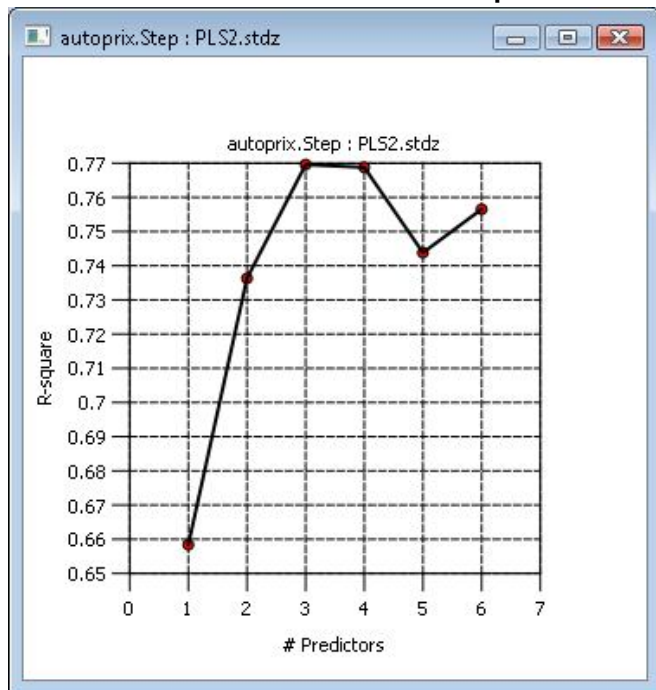


Training R <sup>2</sup>	0.74
CV-R <sup>2</sup>	0.69 (.05)
<b>Standardized</b>	
<b>Predictors</b>	<b>Coefficient</b>
CYLINDER	0.64
WEIGHT	0.23

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
<b>Total</b>	<b>138</b>	12	12	30	12	12	12	12	12	12	12
<b>Predictors</b>		2	2	5	2	2	2	2	2	2	2

# Example: PLS2 with Standardized Predictors

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

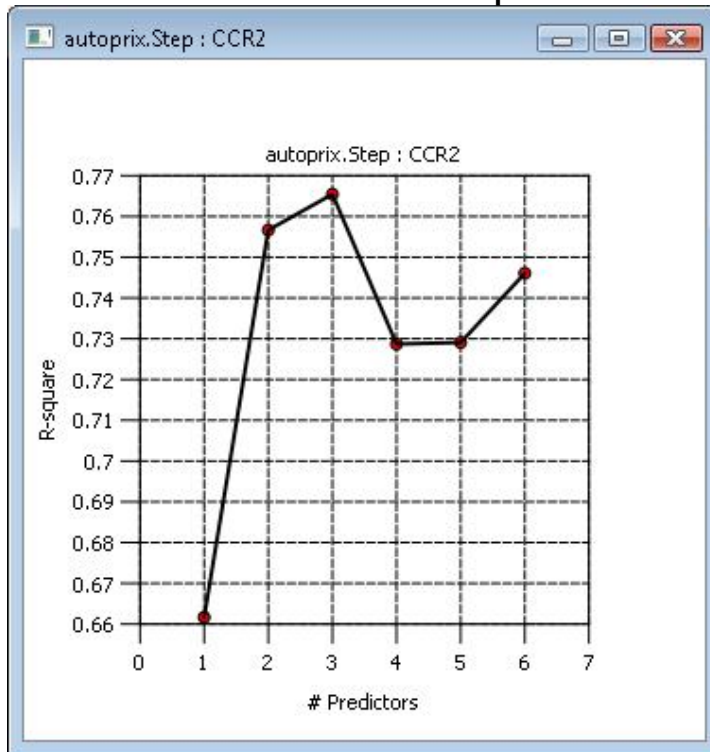


<b>Training R<sup>2</sup></b>	<b>0.84</b>
<b>CV-R<sup>2</sup></b>	<b>0.78 (.02)</b>
<b>Standardized Predictors</b>	<b>Standardized Coefficient</b>
ZPOWER	0.58
ZCYLINDER	0.20
ZWEIGHT	0.19

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: 2-component CCR Model (CCR2)

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

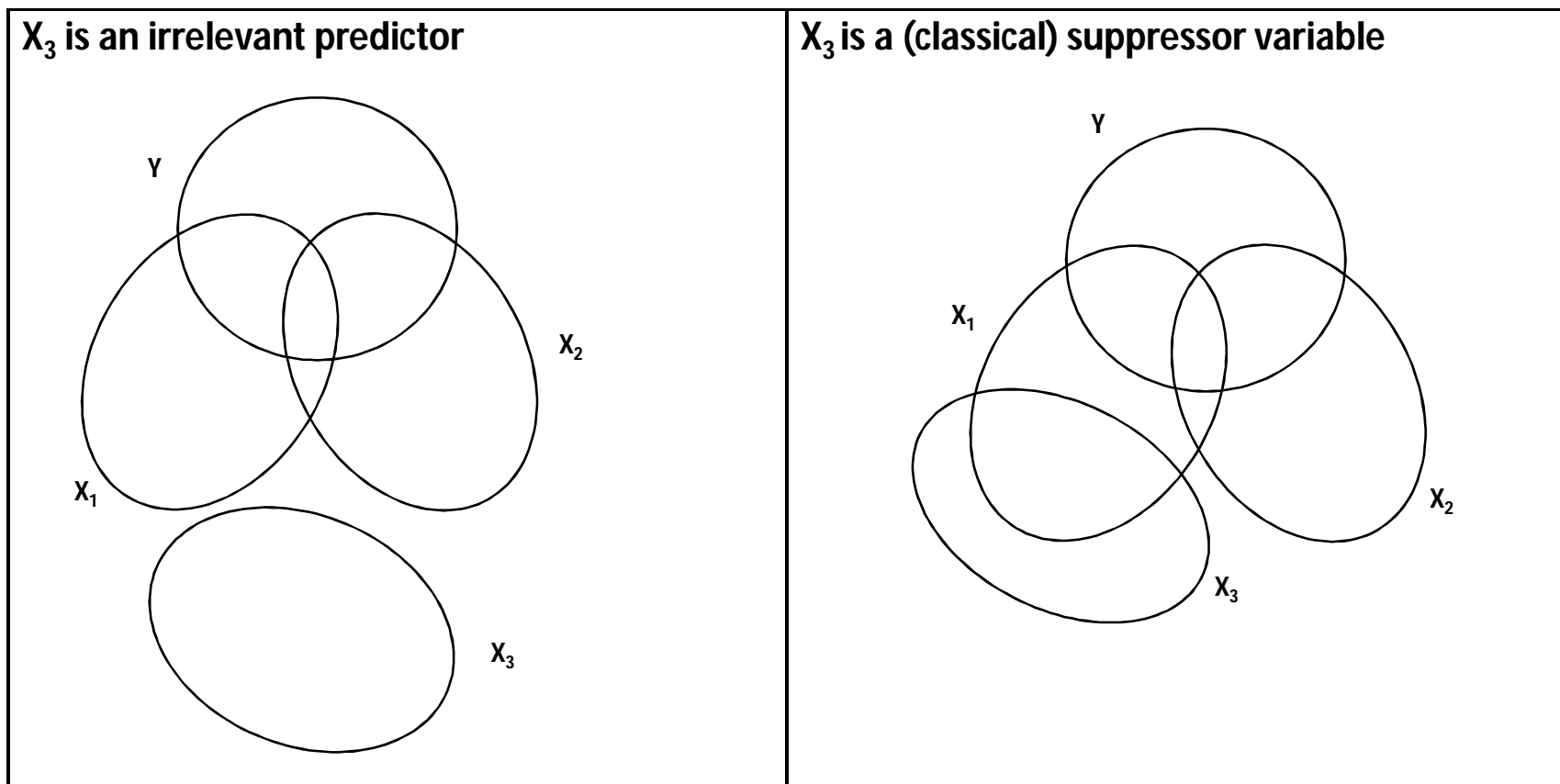


Training R <sup>2</sup>	0.84
CV-R <sup>2</sup>	0.77 (.03)
<b>Predictors</b>	<b>Standardized Coefficient</b>
POWER	0.45
WEIGHT	0.44
SPEED	0.10

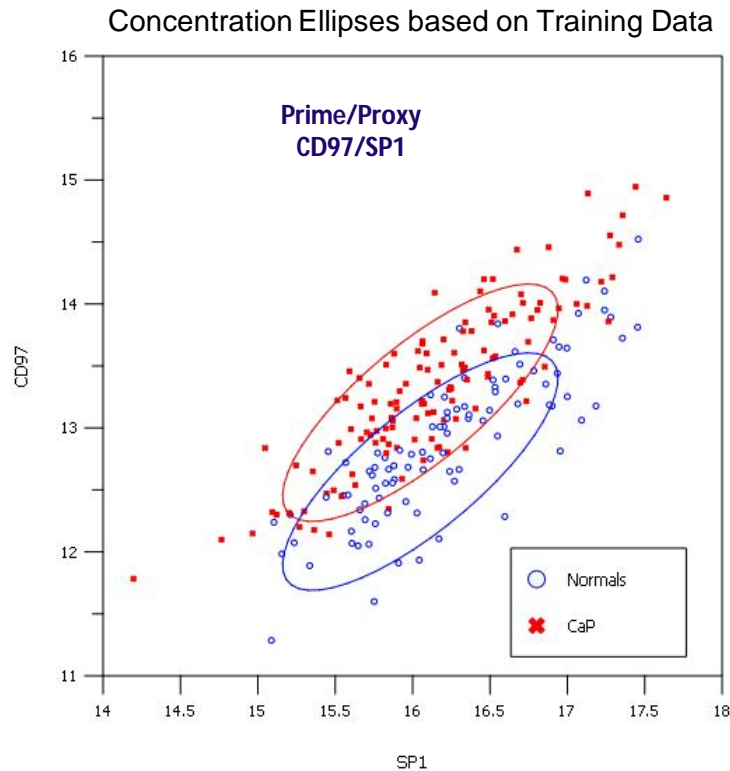
Predictor	All	1	2	3	4	5	6	7	8	9	10
POWER	60	6	6	6	6	6	6	6	6	6	6
WEIGHT	59	6	6	6	6	6	5	6	6	6	6
SPEED	27	3	6	3	3	2	0	3	4	0	3
CYLINDER	23	2	6	3	2	3	1	3	1	0	2
LENGTH	10	1	6	0	0	1	0	0	1	0	1
WIDTH	7	0	6	0	1	0	0	0	0	0	0
Total	186	18	36	18	18	18	12	18	18	12	18
Predictors		3	6	3	3	3	2	3	3	2	3

# 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.

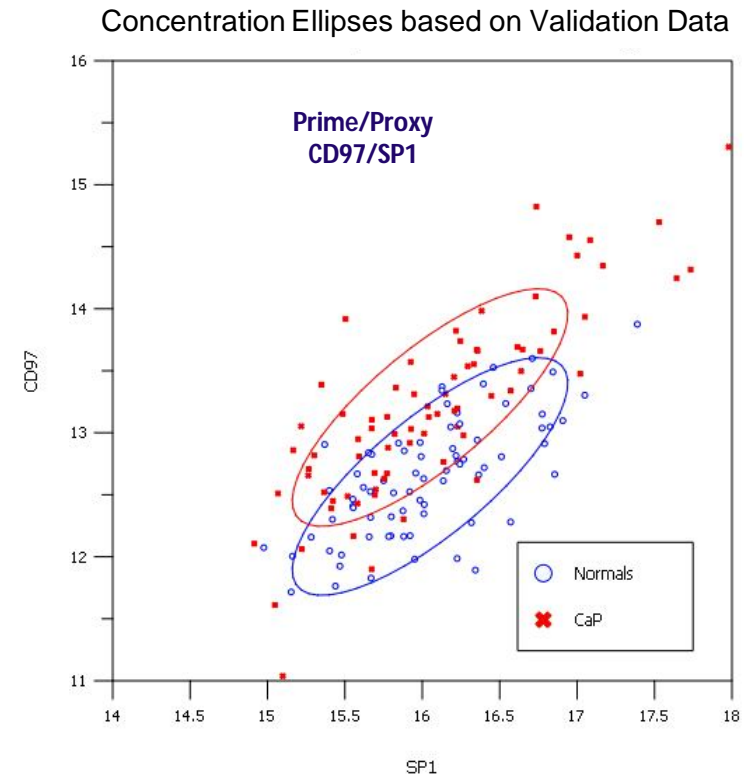


## Example of Suppressor Variable in 2-Gene Model Providing Good Separation of Prostate Cancer (CaP) vs. Normals, Confirmed by Validation Data



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.



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 (LM)

- **14 Valid predictors**, including an important suppressor, SP1
- **42 Extraneous predictors** (i.e., true coefficients equal zero for these)
- 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).

Of the 42 *extraneous* predictors, 14 (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 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.

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

## 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.

# 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
  - ✓ 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

## CCR Variants in CORExpress®

CCR-Logistic: Logistic Regression Models

CCR-LDA: Linear Discriminant Analysis

CCR-Cox: Survival (Event History) Models:

Latent Class Extensions: (under development)

- Separate models for each LC segment
- Selection of predictors prior to LC modeling

# 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:

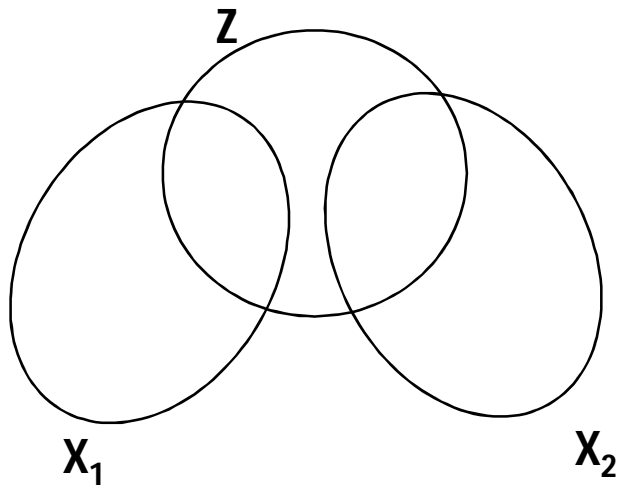
$$\text{Logit}(Z) = \alpha + b_{1.2} S_1 + b_{2.1} S_2$$

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

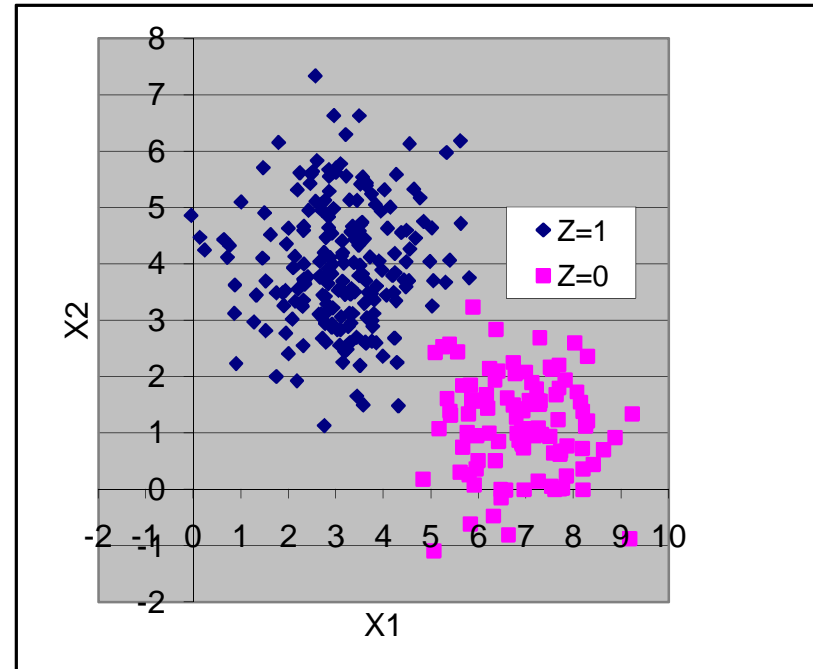
$$\text{Logit}(Z) = \alpha_{.12} + \gamma_{g.1} S_1 + \gamma_{g.2} S_2 + \beta_{g.12} X_g$$

# Naïve Bayes (1-component) Logistic or LDA Model

$X_1$  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



CCR1 (K=1) is equivalent to Naïve Bayes Model

# 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 (LM)**

$G_1 = 14$  preds +  $G_2 = 14$  irrelevant preds correlated with true +  $G_3 = 28$  irrelevant predictors uncorrelated with true;  
Continuous dependent variable,  $N = 50$ , population  $R^2 = .9$ ; 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.

## CCR/Select vs. ISIS for Pre-Screening in Ultra-High Dimensional Data

Fan and Lv (2008) distinguish between high and ultra-high dimensional data. Recognizing the limitations of their SIS screening procedure, they proposed ISIS, an iterative screening method, to pre-screen predictors in ultra-high dimensional data where suppressor variables are present. Fan et. al. (2009) present ISIS simulation results based on 3 prime predictors and one proxy predictor. Our demo data file ISIS400.sav contains such simulated data.

For comparison, we consider the following CCR-based 3-component prescreening step, called **CCR/Select**, to select the best  $M$  predictors, where  $M$  is pre-specified:

For Component 1: Apply Inverse normal transformation to Comp. #1 p-vals  $> .5$  to get  $Zval1$ , and use 2-class truncated normal mixture (latent class) model on  $-Zval1$  to identify the  $G_1$  most significant predictors ( $G_1$  predictors whose posterior prob  $> .5$  of being in class with lowest p-vals). Set component #1 loadings to 0 for all but  $G^*_1$  predictors, where  $G^*_1 = \min\{\max\{G_1, 2\}, 10\}$ .

For Component 2: Compute  $Zval2 =$  Inverse normal of Comp #2 p-vals  $> .5$  (excluding the  $G^*_1$  predictors identified above), and estimate latent class model on  $-Zval2$  to identify  $G_2$  predictors assigned to lowest component #2 p-val class. Set the loading to 0 for all but the  $G^*_2$  predictors with lowest p-values (excluding the  $G^*_1$  predictors), where  $G^*_2 = \min\{\max\{G_2, 1\}, G_1\}$ .

For Component 3: Set the loading to 0 for all but the  $M$  predictors with lowest p-values.

## Results: CCR/Select more often selects all true predictors than ISIS

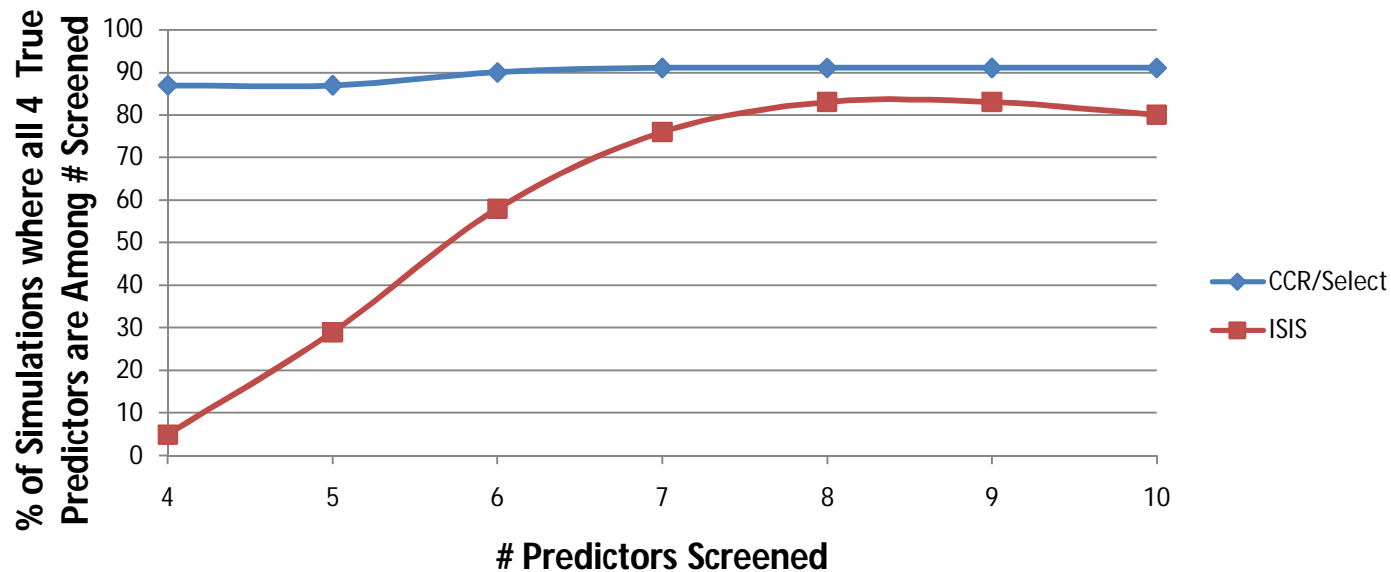
We simulated 100 data sets according to specifications of Fan et. al. (2009) with  $N=200$ :

**Logistic Regression** with  $\beta_0 = 0$ , effects of primes  $\beta_1 = \beta_2 = \beta_3 = 4$ ; effect of suppressor  $\beta_4 = -6\sqrt{2}$  and predictors  $X_5 - X_{1000}$  are irrelevant:  $\beta_5 = \beta_6 = \dots = \beta_{1000} = 0$ .

$$\text{Logit}(Z) = \beta_0 + \sum_{g=1}^{1000} \beta_g X_g$$

where  $X$  follows a multivariate normal distribution with means 0, variances 1 and all correlations = .5 except for  $\text{corr}(X_g, X_4) = 1/\sqrt{2}$  for  $g=1,2,3$ .

**Simulation (N=200) Screening Results: CCR/Select vs. ISIS**



CCR/Select includes  $X_4$  among 10 top predictors 91% of the time compared to only 80% for ISIS.

## Correlated Component Regression (CCR) Tuning Parameters

**In practice, M-fold cross-validation used to optimize CCR with respect to 2 tuning parameters:**

- 1) Number of components  $K^*$ : often  $K^* = 3$  or 4**
- 2) Number of predictors  $P^*$ : As  $P$  is reduced performance usually improves up to a point beyond which performance decays.**

**The ability of CCR to capture the effects of suppressor variables in component #2 is a key to its good performance.**

## Conclusions

When based on *unstandardized* predictors, PLS-R component loadings weight more heavily predictors having higher variance, and therefore may require more components than CCR, in order to reduce the effects of these predictors when such weighting is not warranted. Performing PLS-R with *standardized* predictors yields the same predictions as CCR for  $K = 1$ , but not for  $K > 1$ . Also, use of M-fold CV for PLS-R with *standardized* predictors is not as appropriate, because these standardized predictors no longer have variance 1 when excluding cases in a fold.

CCR loadings and coefficients differ from those of PLS-R in that they are invariant to predictor scale, and CCR components are correlated. These differences tend to yield fewer, more interpretable components than PLS-R.

The CCR step-down algorithm can help improve prediction and interpretation when extraneous or completely irrelevant variables are included among the candidate predictors. With *unstandardized* predictors the step-down algorithm might be expected to work less well for PLS-R than CCR, because the optimal number of components for PLS-R is a function of predictor scale, which makes it more difficult to determine both the optimal number of components and number of predictors simultaneously.

When suppressor variables exist in data, they should be included in predictive models because they can improve prediction substantially. CCR has higher power for capturing effects of suppressor variables than stepwise regression or penalized regression (lasso, ridge regression).

# References

Bair, E., T. Hastie, P. DeBashis, 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 Mutlicollinearity 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.