



Correlated Component Regression: An Alternative to PLS-regression and Other Regression Approaches in Situations with Many Correlated Predictors

**Jay Magidson
Statistical Innovations Inc.
Belmont, MA. USA**

Outline of Presentation

- Regression problem with many correlated predictors
- Importance of including suppressor variables among predictors
- Correlated Component Regression* (CCR)
 - Without variable selection – 1 tuning parameter K
 - With variable selection – 2 tuning parameters (K, P)
- Comparisons with other sparse regression methods:
 - Penalized regression -- lasso, Elastic Net
 - Sparse PLS Regression
- Summary and some extensions

*CCR is implemented in the CORExpress™ program (patent pending).

Regression Problem with Many Correlated predictors

Problem:

In exploratory regression when the number of correlated predictor variables P approaches or exceeds sample size N , coefficients estimated with traditional techniques become unstable or are not unique due to multicollinearity (singularity of the covariance matrix).

Approaches for dealing with these problems include:

- **Penalized regression approaches – lasso, elastic net**
- **Dimension reduction approaches – predict based on $K^* < \min(P, N-1)$ components**
 - **PLS Regression (PLS-R) – components are orthogonal**
 - **Correlated Component Regression* (CCR) – components are correlated to accommodate suppressor variables**

To avoid over-fitting due to extraneous predictors, methods producing *sparse* solutions that include only $P^* < P$ predictors are of particular interest.

Importance of Including Suppressor Variables among Predictors

Suppressors are predictors that are uncorrelated with the dependent variable

When included in model, a suppressor improves prediction by enhancing the effects of one or more other predictors in the model.

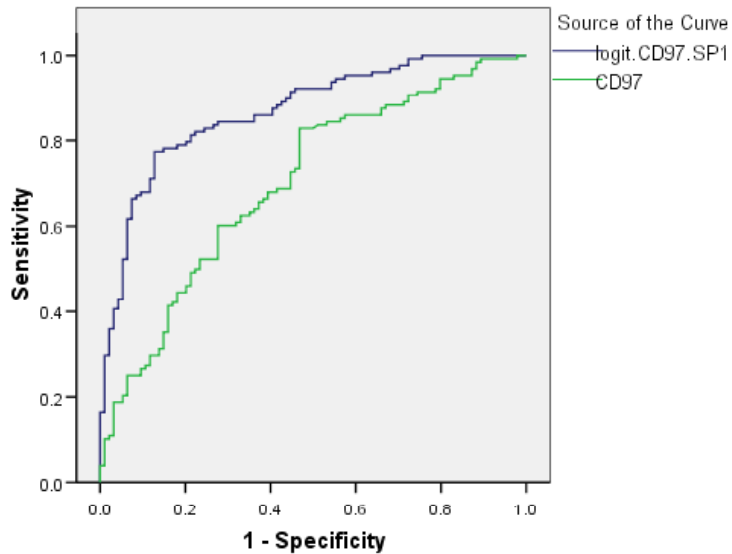
Often, suppressors are among the most important predictors in the model (Magidson and Wassmann, 2010).

The importance of suppressor variables are well documented in the statistics literature (e.g., see Horst, 1941; Lynn, 2003; Friedman and Wall, 2005).

Example of Suppressor: Adding SP1 as 2nd Predictor Significantly Improves Classification of Prostate Cancer (CaP) vs. Normal Subjects Over 1-Gene Model based on CD97 Only

Enhancer Effect of Proxy Gene

Training



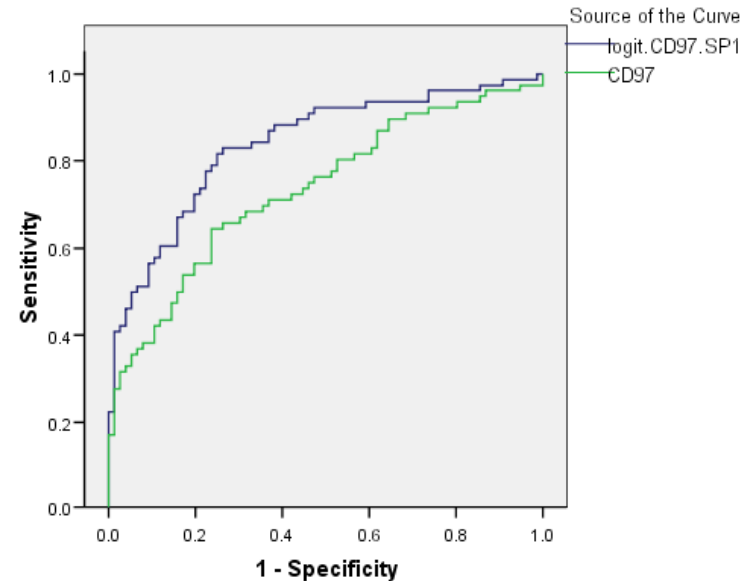
Diagonal segments are produced by ties.

| | | |
|---|----------------|---------|
| 2-gene model: CD97 + SP1 | AUC = | 0.87 |
| 1-gene model: CD97 | AUC = | 0.70 |
| Improvement in AUC by Inclusion of SP1 | Δ AUC = | 0.17 |
| AUC Difference p-value | p-val = | 7.6E-05 |
| Logit Model Unique Contribution for SP1 | p-val = | 1.4E-11 |

| | | |
|--------------------------------------|---------|------|
| 1-gene model: SP1 (no direct effect) | AUC = | 0.52 |
| AUC p-value | p-val = | 0.57 |

Training Set Results:
 CaP (N=128) vs. Normals (N=94)

Validation



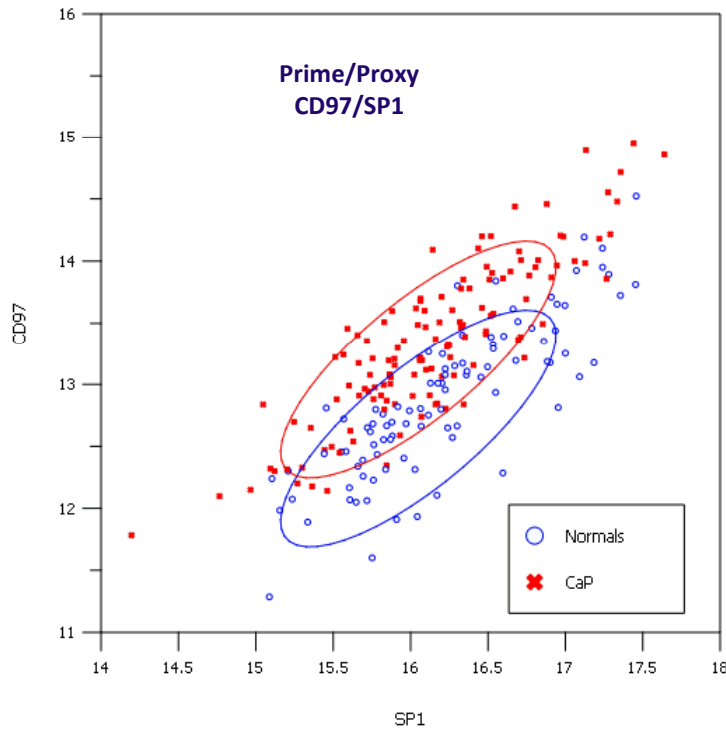
| | | |
|---|----------------|---------|
| 2-gene model: CD97 + SP1 | AUC = | 0.84 |
| 1-gene model: CD97 | AUC = | 0.73 |
| Improvement in AUC by Inclusion of SP1 | Δ AUC = | 0.10 |
| AUC Difference p-value | p-val = | 0.047 |
| Logit Model Unique Contribution for SP1 | p-val = | 2.2E-06 |

| | | |
|-------------------|---------|------|
| 1-gene model: SP1 | AUC = | 0.50 |
| AUC p-value | p-val = | 0.93 |

Validation Set Results:
 CaP (N=76) vs. Normals (N=76)

Graphical Explanation of Suppressor Effect in Logistic Regression

Concentration Ellipses based on Training Data

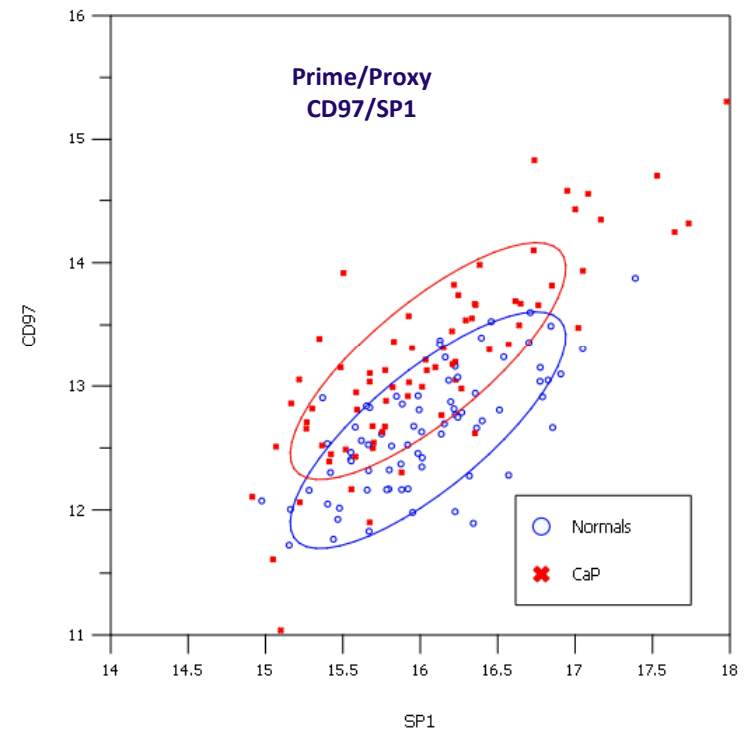


CaP Subjects have higher CD97 levels as compared to Normals – Red ellipse lies above blue ellipse.

Since CaP and Normals do not differ on SP1, SP1 has no direct effect. It improves classification by removing irrelevant variation from CD97.

Note: These 2 genes are highly correlated.

Concentration Ellipses based on Validation Data



The validation data supports the hypothesis 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.

Shown in Delta CT units

How Suppressor Variable SP1 Enhances the Effect of CD97

Magidson and Wassmann (2010) refer to a gene that acts as a suppressor variable as a *proxy gene*, enhancing the effect of the associated correlated *prime gene* by removing irrelevant variation.

SP1 functions solely as a proxy gene to enhance the effect of the Prime gene CD97:

$$\begin{aligned} \text{Logit} &= \alpha + \beta_1 * [CD97 - (a + b * SP1)] \\ &= \alpha' + \beta_1 * CD97 - \beta_2 * SP1 \end{aligned}$$

$$\beta_2 = b * \beta_1 \quad \textit{Equivalence due to SP1 being a pure proxy gene}$$

where $a + b * SP1$ is the OLS prediction of CD97 by SP1

Suppressor variables (proxy genes) commonly occur in practice and are among the most important predictors.

Biological explanation: SP1 is a transcription factor that has been extensively studied for its role in regulation of gene expression for many genes. Specifically, SP1 is known to regulate CD97.

Correlated Component Regression

Correlated Component Regression (CCR) is a general approach for the development of a sequential K-component predictive model, each component estimated by application of the naïve Bayes rule to handle multi-collinearity.

- The first component S_1 captures the effects of the predictors which have direct effects ("prime predictors"). 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 ("proxy predictors") that improve prediction by removing extraneous variation from one or more *prime predictors*.
- Additional components are included if they improve prediction significantly.

Prime predictors are identified as those having large loadings on S_1 , and *proxy predictors* as those having large loadings on S_2 , and small or zero loadings on component #1.

- Variable reduction may be achieved using a step-down algorithm which at each step the least important predictor (or the least important 1% of predictors) is removed, importance defined by the absolute value of the standardized coefficient.

Example: Correlated Component Regression / Linear Regression

Step 1: Form 1st component S_1 as weighted average of all 1-predictor models (ignoring α_g)

$$\hat{Y}_g = \alpha_g + \beta_g X_g \quad g=1,2,\dots,P; \quad S_1 = \frac{1}{P} \sum_{g=1}^P \beta_g X_g$$

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

Step 2: Form 2nd component S_2 as weighted average of $\beta_{g.1} X_g$ where each $\beta_{g.1}$ is estimated from the following 2-predictor model

$$\hat{Y}_{g.1} = \alpha_{.1} + \gamma_g S_1 + \beta_{g.1} X_g \quad g=1,2,\dots,P; \quad S_2 = \frac{1}{P} \sum_{g=1}^P \beta_{g.1} X_g$$

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

$$\hat{Y}(2) = \alpha + b_{1.2} S_1 + b_{2.1} S_2$$

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

$$\hat{Y}_{g.12} = \alpha_{.12} + \gamma_{g.1} S_1 + \gamma_{g.2} S_2 + \beta_{g.12} X_g$$

Example: Correlated Component Regression / Logistic Regression

Step 1: Form 1st component S_1 as weighted sum of P 1-predictor models (ignoring α_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 weighted sum of $\beta_{g.1} X_g$
Where each $\beta_{g.1}$ is estimated from the following $P=2$ -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:

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

Example: Correlated Component Regression / Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) provides maximum likelihood estimates for the coefficients of logistic regression under the following assumptions:

- a) X_s follow MVN distribution in each group $Z=1$ and $Z=0$
- b) Variance-Covariance matrix is equal within each group

With high dimensional data (small N and large P) it has been shown that use of the 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)

even when the true model is that of LDA!

1-component CCR is equivalent to Naïve Bayes.

For $P \leq N$, *saturated CCR** is equivalent to traditional regression – linear, logistic, LDA, etc.)

Simulation results show that K -component CCR, with $K \in [2-6]$, typically produces best *outside-the-sample* prediction.

* For *saturated CCR*, $K = \min(P, N-1)$: Optimal K^* chosen by M -fold cross-validation

M-fold Cross-Validation (Without Variable Selection) Determines Tuning Parameter K

Divide sample into M equal groups (folds).

For specified # components K, estimate model M times, each time omitting one fold.

Compute error or other performance criterion (e.g., AUC) based on cases in omitted fold.

Average performance over all M omitted folds.

Choose K^* that performs best.

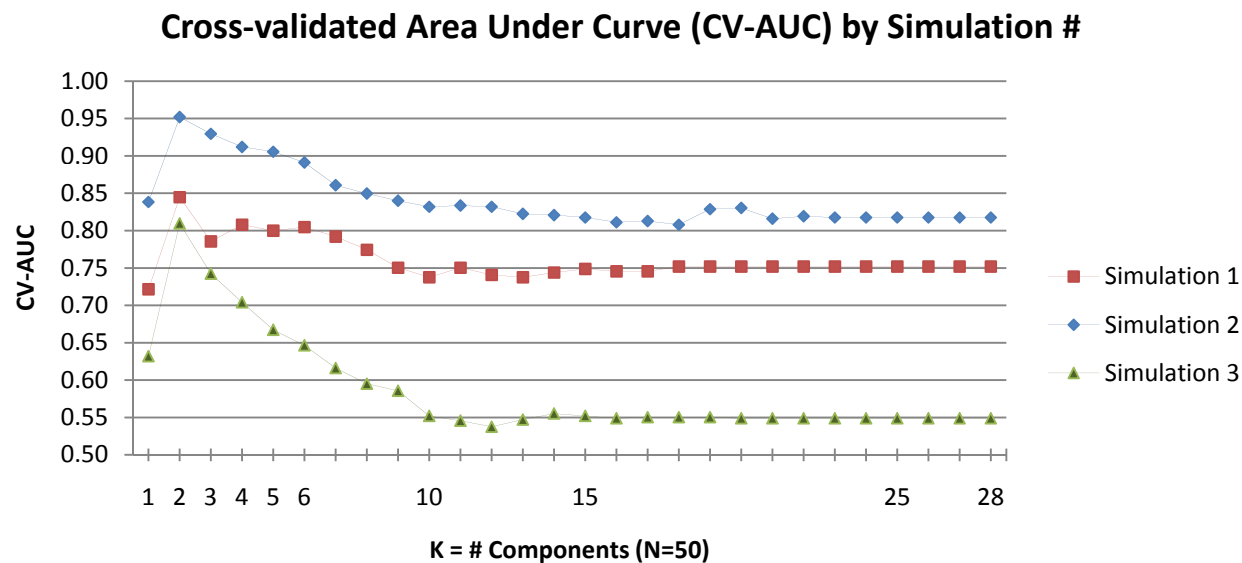
CCR($K \in [2-6]$) Outperforms Both Naïve Bayes (CCR: $K = 1$) and LDA (CCR: $K = P = 28$)

Data Simulated Under LDA Assumptions with moderate number of predictors ($P = 28$)

- 8 stronger predictors + 20 weaker predictors, all moderately correlated
- $N = 50$: $N_1 = N_2 = 25$

Results from 5-fold cross-validation: 2-component CCR model tends to be best ($K^* = 2$)

- Even here Naïve Bayes ($K=1$) outperforms LDA ($K=28$) in 2 of 3 simulations
- CCR with $K \in [2,6]$ substantially outperforms LDA and Naïve Bayes



Correlated Component Regression with Step-down Variable Selection (Reduction)

Step-down: For a given K^* -component model, eliminate the variable that is the least important, where importance is quantified by the absolute value of the variable's standardized coefficient, where the standardized coefficient is defined as:

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

For example, suppose that comparing the absolute value of the standardized coefficients for the K^* -component model determines that predictor g^* is the least important. Then that predictor would be excluded and the steps of the CCR estimation algorithm are repeated on the reduced set of predictors.

When P falls below K^* , set $K^* = P$ and continue estimating saturated models. Thus, step-down can always continue to $P = 1$ predictor

M-fold Cross-Validation with Step-Down Variable Reduction – 2 Tuning Parameters

Divide sample into M equal groups (folds).

For specified # components K , estimate model **with step-down** M times, each time omitting one fold.

Compute performance based on cases in omitted fold.

Compute average performance over all M omitted folds.

Estimate $CCR(K^*, P^*)$ based on full sample.

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

High-Dimensional Data: CCR with $P > N$: $P = 28 + 28 + 28 = 84$; $N = 50$

Data Simulated Under LDA Assumptions with $P=84$ predictors

- Original 28 “true” predictors with variance-covariance matrix Σ_{28}
- Additional 28 extraneous predictors, uncorrelated with 28 true predictors but maintaining same variances and covariances Σ_{28}
- Additional 28 irrelevant predictors, each uncorrelated with all other predictors

$$\Sigma = \begin{pmatrix} \Sigma_{28} & | & \mathbf{0}_{28} & | & \mathbf{0}_{28} \\ \mathbf{0}_{28} & | & \Sigma_{28} & | & \mathbf{0}_{28} \\ \mathbf{0}_{28} & | & \mathbf{0}_{28} & | & \mathbf{0}_{28} \end{pmatrix}$$

When all $P = 84$ predictors are included in regression, CV-step-AUC = .75 ($K^* = 3$)

Substantial improvement occurs with $P^* = 6$ predictors, CV-step-AUC = .85 ($K^* = 6$)

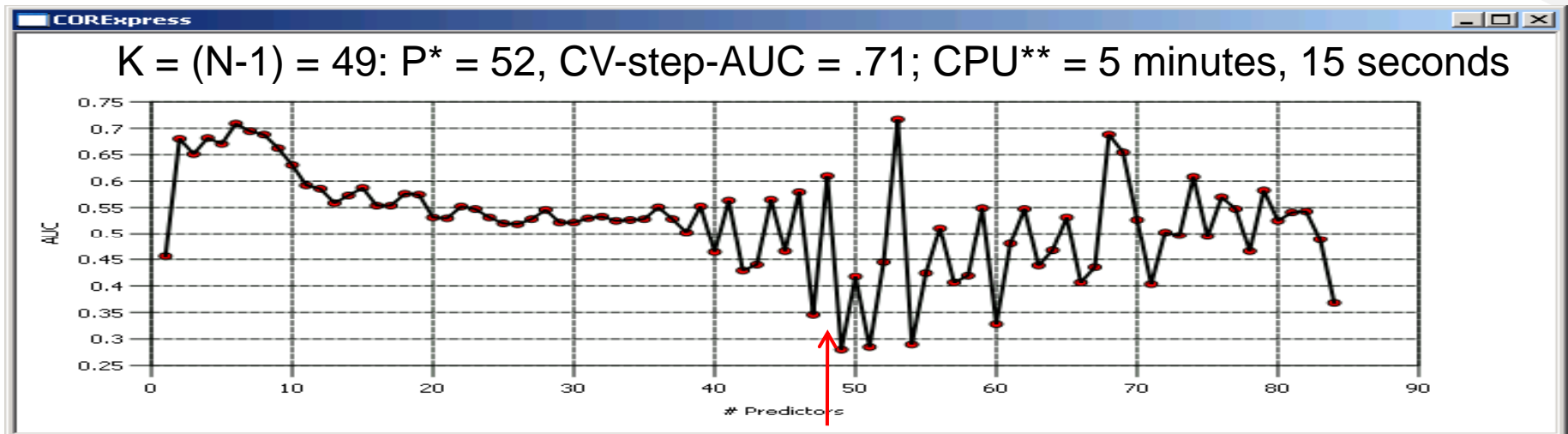
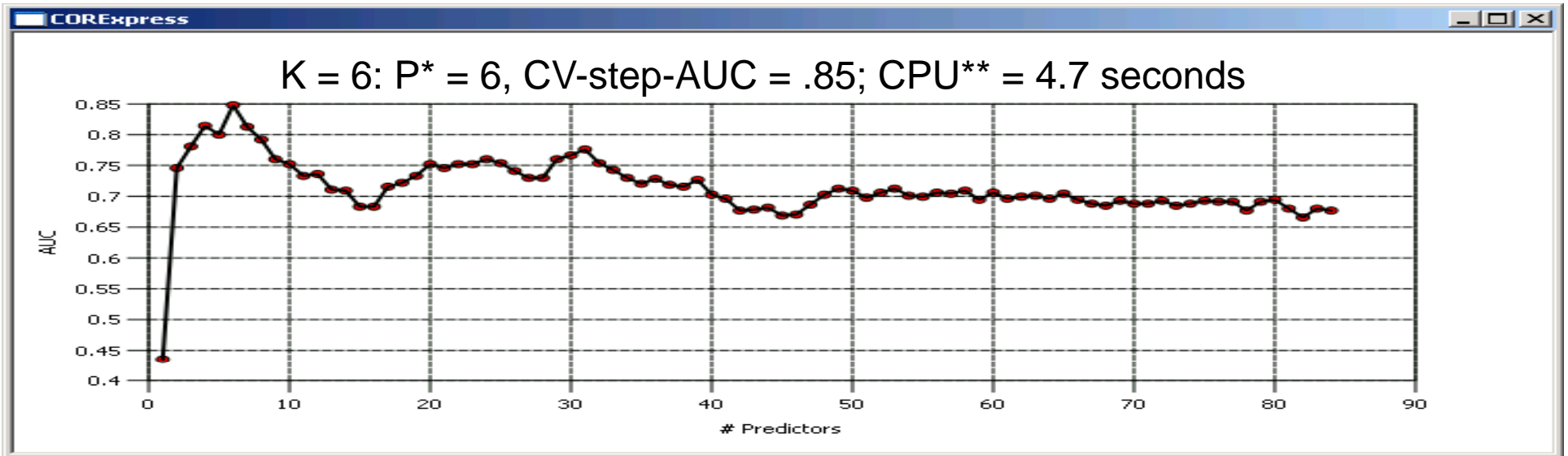
Note: CCR(6, 6) is an LDA model with 6 selected predictors

Results from simulated sample #1
Comparison of AUC for $P = 84, 53, 6$

| K P | CV-AUC | | | P-optimal |
|-------|--------|------|------|------------------|
| | 84 | 53 | 6 | |
| 1 | 0.64 | 0.62 | 0.69 | .84 ($P^*=2$) |
| 2 | 0.69 | 0.66 | 0.70 | .75 ($P^*=1$) |
| 3 | 0.75 | 0.75 | 0.68 | .78 ($P^*=2$) |
| 4 | 0.70 | 0.72 | 0.79 | .79 ($P^*=6$) |
| 5 | 0.69 | 0.68 | 0.82 | .82 ($P^*=6$) |
| 6 | 0.68 | 0.72 | 0.85 | .85 ($P^*=6$) |
| 7 | 0.70 | 0.65 | 0.82 | .83 ($P^*=7$) |
| 49 | 0.37 | 0.72 | 0.71 | .72 ($P^*=52$) |

Sample #1: Resulting Model CCR (6, 6) Outperforms Saturated CCR Model

CV-step-AUC = .85 vs. .71



**CPU times based on CORExpress™ program run on 2Ghz CPU with XP

Problem: Many Variable Pre-Screening Methods Exclude Suppressor Variables

Problem and solution:

For ultra-high dimensional data with many irrelevant predictors, typical with gene expression data, by chance some large loadings for the many irrelevant predictors may dominate the first component, leading to unreliable results. To avoid this, an initial variable selection 'screening' step may be performed to reduce # genes to a manageable number prior to model estimation.

Most current screening methods typically exclude the important proxy genes (thus, suboptimal) – e.g., supervised principal components analysis/SPCA: Bair, et. al. , 2006; SIS: Fan and Lv, 2008.

Fan. et. al (2008, 2009) propose ISIS, an iterative screening method designed to remedy the omission of such predictors by SIS, and shows the improvement over SIS with simulated data. However, ISIS has been criticized for having too many tuning parameters. We are developing a CCR-based screening procedure, CCR/Select, that has a single parameter M , or the desired number of predictors to be selected (Magidson and Yuan, 2010).

The next slides introduce CCR/Select and compare its performance with ISIS based on Fan et. al. (2009) simulated data.

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

Fan and Lv (2008) distinguish between high and ultra-high dimensional data, and propose **ISIS** 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.

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 Z_{val1} , and use 2-class truncated normal mixture (latent class) model on $-Z_{val1}$ 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 Z_{val2} = Inverse normal of Comp #2 p-vals $> .5$ (excluding the G^*_1 predictors identified above), and estimate latent class model on $-Z_{val2}$ 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

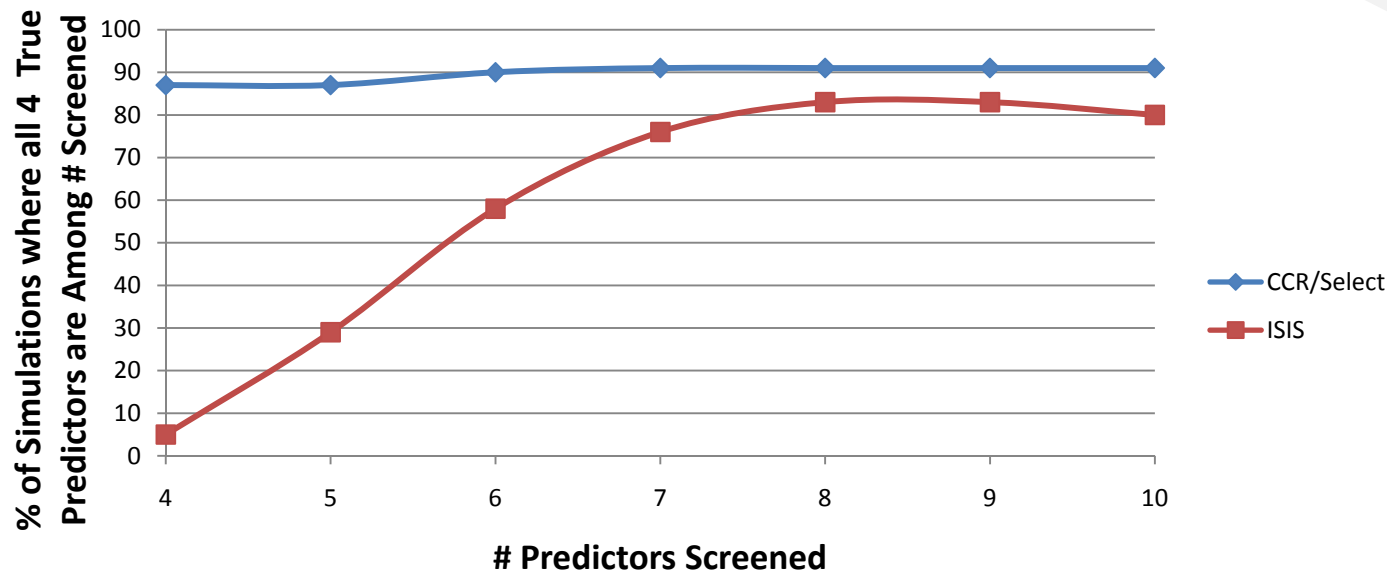
We simulated 100 data sets according to specifications of Fan et. al. (2009) with N=200:

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.

Simulation Results with Sparse Continuous Predictors based on Assumptions of Linear Discriminant Analysis

Design: Data simulated according to assumptions of Linear Discriminant Analysis

$G_1 = 28$ 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 same sized validation file. Final models from each method evaluated based on large independent 'test' file.

Variable selection 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)

Results favor CCR over the other approaches (Magidson and Yuan, 2010)

Lowest misclassification error rate:

CCR (17.4%), sparse PLS (19.1%), Elastic net (20.2%), lasso (20.8%)

Fewest irrelevant variables:

CCR (3.4), lasso (6.2), Elastic net (11.5), sparse PLS (13.1)

Most sparse solution (average # predictors in model):

CCR (14.5), lasso (17.3), Elastic net (28.3), sparse PLS (32.3)

Conclusions

When suppressor variables exist in data, they should be included in predictive models because they can improve prediction substantially.

CCR has outperformed various penalty approaches as well as PLS regression algorithms in our analyses conducted on high-dimensional simulated and real data based on linear, logistic, and Cox-type survival models, as well as linear discriminant-type models to date. All data sets we have used contain at least one suppressor variable.

In the case of ultra-high dimensional data, a variable pre-screening step may be needed. Many current variable selection algorithms should be avoided as they are designed to select only predictor variables that are correlated with the dependent variable and thus exclude suppressor variables. We are currently exploring the use of a CCR-based screening method, and comparing its performance with ISIS. Preliminary results suggest that a CCR-based screening method may improve over ISIS in certain settings.

Correlated Component Regression (CCR) is a Promising New Regression Method

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.
- Chun, H. and Keles, S. (2009). "Sparse partial least squares for simultaneous dimension reduction and variable selection". *Journal of the Royal Statistical Society - Series B*.
- Fort, G. and Lambert-Lacroix, S. (2003). Classification Using Partial Least Squares with Penalized Logistic Regression. IAP-Statistics, TR0331.
- 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.
- 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.
- Hanczar, Zucker, J., Henegar, C. and L. Saitta (2007). Feature Construction from Synergic Pairs to Improve Microarray-based Classification. *Bioinformatics*. Vol. 23, No. 21 2007, pages 2866-2872.
- 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.
- Lynn, H (2003). Suppression and Confounding in Action. *The American Statistician*, Vol.57, 2003.
- Magidson, J. (2010). A Fast Parsimonious Maximum Likelihood Approach for Prediction Outcome Variables from a Large Number of Predictors. Presentation at COMPSTAT 2010 meetings.
- 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. (2010) "Correlated Component Regression: A Prediction/Classification Methodology for Possibly Many Features", Proceedings of the American Statistical Association.
- Magidson, J. and Y. Yuan (2010) "Comparison of Results of Various Methods for Sparse Discriminant Analysis", unpublished report #CCR2010.1, Belmont MA: Statistical Innovations.
- Zou, H. and T. Hastie (2005). "Regularization and variable selection via the elastic net". *J.R. Statist. Soc. B.* 67, Part 2, pp.301-320.





Thank you for your attention!