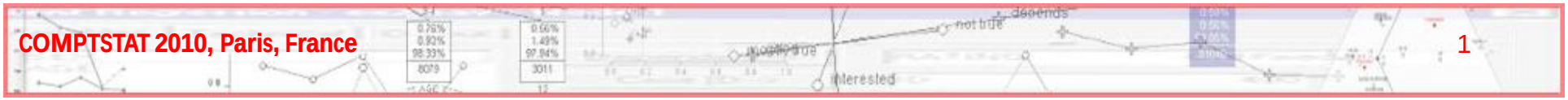
Statistical Innovations

Thinking outside the brackets!™

# Correlated Component Regression: A Fast Parsimonious Approach for Predicting Outcome Variables from a Large Number of Predictors

Jay Magidson, Ph.D.
Statistical Innovations



# Correlated Component Regression (CCR)

New methods are presented that extend traditional regression modeling to apply to high dimensional data where the number of predictors P exceeds the number of cases N (P >> N). The general approach yields K correlated components, weights associated with the first component providing direct effects for the predictors, and each additional component providing improved prediction by including suppressor variables and otherwise updating effect estimates. The proposed approach, called Correlated Component Regression (CCR), involves sequential application of the Naïve Bayes rule.

With high dimensional data (small samples and many predictors) 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)

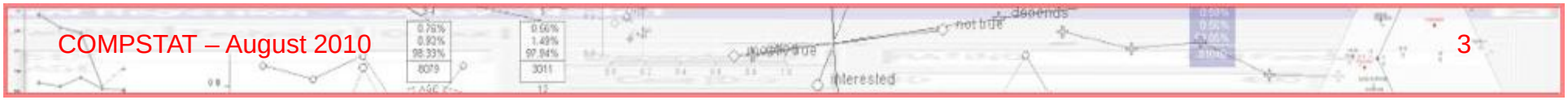
even when the true model is that of LDA! Results from simulated and real data suggest that CCR outperforms other sparse regression methods, with generally good *outside-the-sample* prediction attainable with K=2, 3, or 4.

When P is *very* large, an initial CCR-based variable selection step is also proposed.

# Outline of Presentation

- The P > N Problem in Regression Modeling
- Important Consideration: Inclusion of Suppressor Variables
- Sparse Regression Methods
  - Penalty approaches -- lasso, Elastic Net (GLMNET)
  - PLS Regression (PLSGENOMICS, SPLS)
  - Correlated Component Regression (CORExpress™)
- Results from Simulations and Analyses of Real Data

- Initial Pre-screening Step for Ultra-High Dimensional Data
- Planned Correlated Component Regression (CCR) Extensions



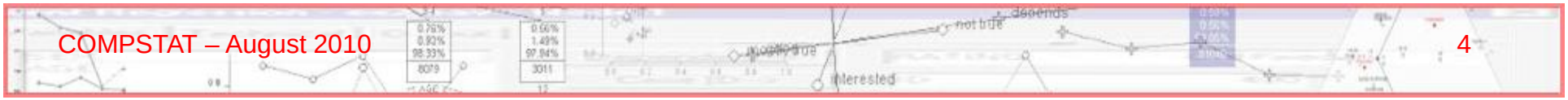
# The P > N Problem in Regression Modeling

**Problem 1:**
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 multicolinearity (singularity of the covariance matrix), and in logistic regression, perfect separation of groups occurs in the analysis sample. The apparent good performance often is due to *overfitting, and* will not generalize to the population, performing worse than more parsimonious models when applied to new cases outside the sample.

Approaches for obtaining more parsimonious (or regularized) models include:

- **Penalty methods – impose *explicit* penalty**
- **Component approaches – exclude higher dimensions**

In this presentation we focus on linear discriminant analysis, and on linear, logistic and Cox regression modeling in the presence of high-dimensional data.



# Example: Logistic Regression with More Features than Cases: P > N

Logistic Regression model for dichotomous dependent variable Z and P predictors:

$$Logit(Z) = \alpha + \sum_{g=1}^{P} \beta_g X_g$$

- As P approaches the sample size N, overfitting tends to dominate and estimates for the regression coefficients become unstable
- Complete separation always attainable for P = N - 1
- Traditional algorithms do not work for P > N as coefficients are not identifiable

# Important Consideration: Inclusion of Suppressor Variables

**Problem 2:**
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". Based on experience with gene expression and other high dimensional data, suppressor variables often turn out to be among the most important predictors:

✓ 6-gene model for prostate cancer (single most important gene, SP1, is a proxy gene)
✓ Survival model for prostate cancer (3 prime and 3 proxy genes supported in blind validation)
✓ Survival model for melanoma (2 proxy genes in 4-gene model supported in blind validation)

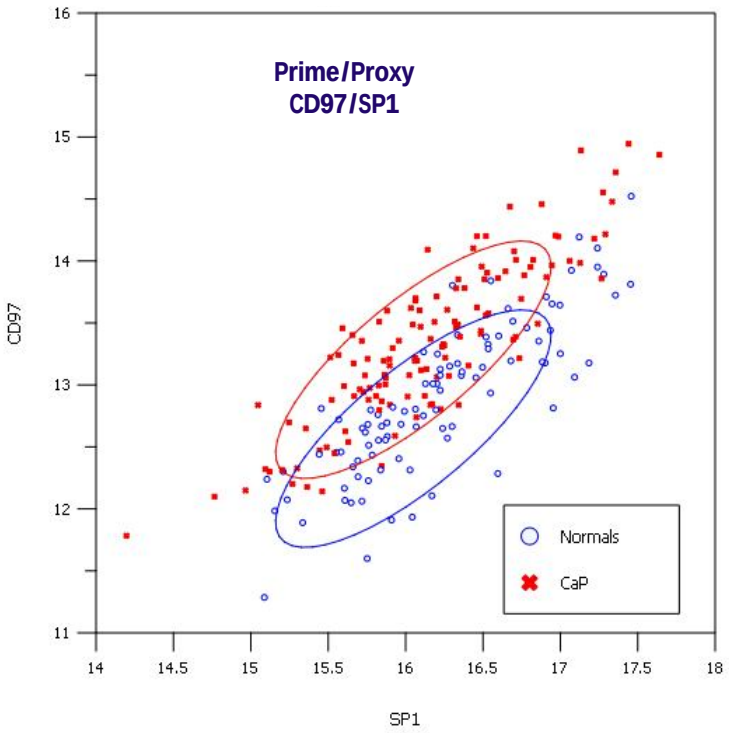
Despite the extensive literature documenting the strong enhancement effects of suppressor variables (e.g., Horst, 1941, Lynn, 2003, Friedman and Wall, 2005), **most pre-screening methods omit proxy genes prior to model development, resulting in suboptimal models.**

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

Because of their sizable correlations with associated prime genes, proxy genes can also provide structural information useful in assuring that these associated prime genes are selected with the proxy gene(s), improving over non-structural penalty approaches.

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

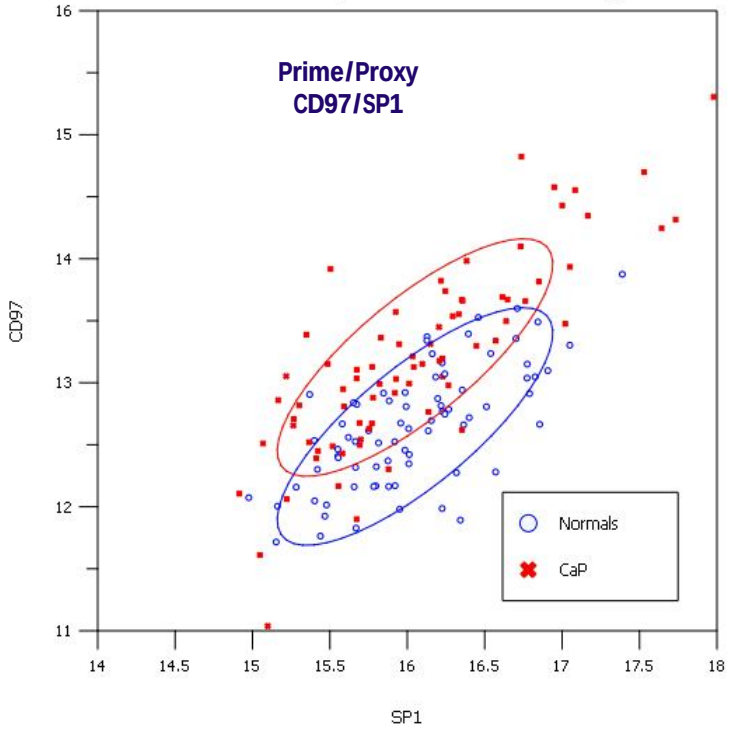
Concentration Ellipses based on Training Data




CaP Subjects have elevated CD97 Δct level as compared to Normals – Red ellipse lies above blue ellipse.

CaP and Normals do not differ on SP1, despite its high correlation with CD97.

Concentration Ellipses based on Validation Data




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

# Some Sparse Regression Approaches

*Sparse* means method involves simultaneous regularization and 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)

C) CCR: Correlated Component Regression – designed to include suppressor variables
P predictors replaced by K < P *correlated components* each defined as a linear combination of the P (or a subset of the P) predictors: CORExpress™ program
  -- Magidson (2010)

# Correlated Component Regression Approach*

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 *prime 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 (*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 significant loadings on $S_1$, and *proxy predictors* as those having significant loadings on $S_2$, and non-significant loadings on component #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. K-fold cross-validation is used to determine the number of components and predictors.

*Multiple patent applications are pending regarding this technology

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

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

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

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

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:

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

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

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

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

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

# Other CCR Variants

**1) Linear Discriminant Analysis: CCR-LDA**

Utilize the random X normality assumption to speed up algorithm.

In step K, *regress each predictor on Z*, controlling for $S_1,\ldots,S_{K-1}$ in fast *linear* regressions:

e.g., for K=1:

$$X_g = \alpha_g + \beta_g' Z \qquad g=1,2,\ldots,P;$$

$$S_1 = \frac{1}{P}\sum_{g=1}^{P} \beta_g X_g$$

$$\beta_g = \beta_g' / MSE$$

where $\beta_g$ is maximum likelihood estimate for log-odds ratio in simple logistic regression model (Lyles et. al., 2009)

**2) Ordinal Logistic Regression: CCR-Logist, CCR-LDA** (extended to ordinal dependent)
For ordinal, Z categories takes on numeric scores (Magidson, 1996)

**3) Survival Analysis: CCR-Cox** – Model expressed as Poisson Regressions (Vermunt, 2009)

**4) Linear Regression: CCR-LM** – for improved efficiency, in step K each predictor is regressed on Z (single application of multivariate linear regression, controlling for $S_1,\ldots,S_{K-1}$)

# Correlated Component Regression Step-down Variable Reduction Step

**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 the loadings associated with the 1st and 2nd components are statistically significant, but those associated with the 3rd component are not. Then K = 2.

Comparing the absolute value of the standardized coefficients for the K*=2-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.

# CCR-LDA Simulation Results with Many Continuous Predictors

**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 validation data with $N_1 = N_2 = 25$. Final models from each method evaluated based on large independent 'test' file.

**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)

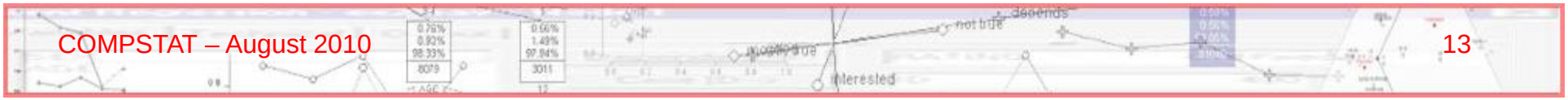
**Results favor CCR over the other approaches** (Magidson and Yuan, 2010)

**Lowest misclassification error rate**:
**CCR (17.4%)**, sparse PLS (19.3%), Elastic Net (21.1%), lasso (21.6%)

**Fewest irrelevant variables**:
**CCR (3.4, 23%)**, lasso (4.3, 31%), Elastic Net (6.6, 34%), sparse PLS (6.9, 34%)

**Most likely to include suppressor variable** (% of simulations):
**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)



# CCR-LM Simulation Results with Many Continuous Predictors

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

$G_1$ = 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.

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

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

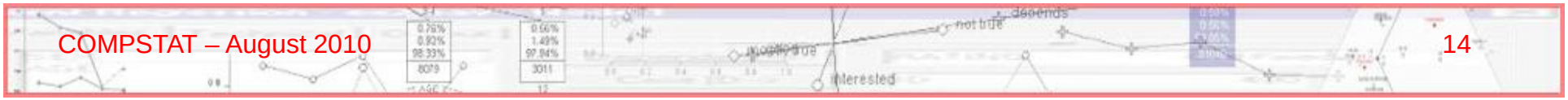
**Number of 'True' Predictors included, Percentage of included that were 'True':**
**CCR (**9.7, **78%)**, TLP (10.3, 50%), sparse PLS (9.5, 48%), Elastic Net (12, 35%)

**Fewest irrelevant *uncorrelated* variables**:
**CCR (1.0, 8%)**, TLP (6.4, 31%), sparse PLS (6.4, 33%), Elastic Net (14.1, 41%)

**Fewest irrelevant *correlated* variables**:
**CCR (1.8, 15%)**, sparse PLS (4.4, 22%), Elastic Net (8.0, 23%), TLP (4.0, 27%)

**Lowest mean squared error**:
**CCR (3.13)**, sparse PLS (3.34), Elastic Net (3.50), TLP (3.55)

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



# Need for Variable Pre-Screening with Ultra-High Dimensional Data

**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 should be avoided because they typically exclude the important proxy genes
– e.g., supervised principle 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 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 $G^*_2 = \min\{\max\{ G_2, 1\}, G_1\}$.

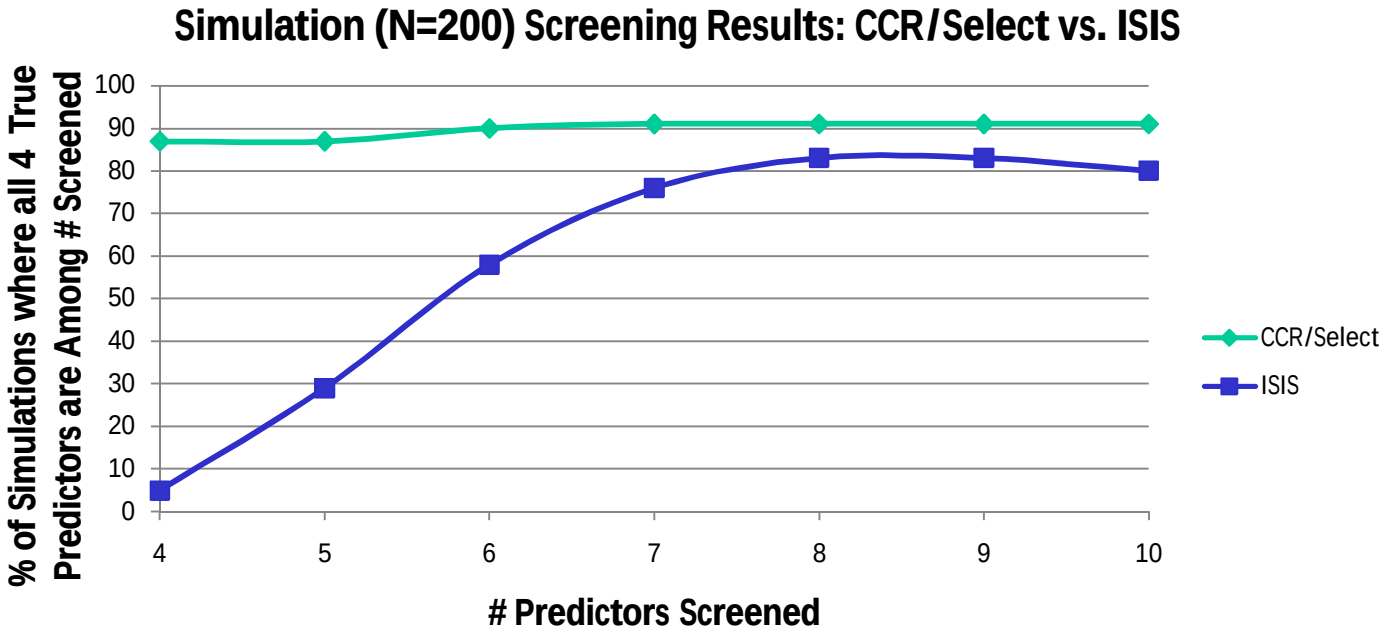
For Component 3: Set the loading to 0 for all but the M predictors with lowest p-values.

# 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: $\beta_5 = \beta_6 = \ldots = \beta_{1000} = 0$.

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

where X follows a multivariate normal distribution with means 0, variances 1 and all correlations = .5 except for $corr(X_i, X_4) = 1/\sqrt{2}$ for $i \neq 4$.




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

# 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

# CCR Variants and Planned Extensions in COREXpress™

CCR-LM: Linear Regression -- Extension to multiple outcome variables planned

CCR-LDA: 2-group Linear Discriminant Analysis -- Extension beyond 2 groups planned

CCR-Logist: Dichotomous and Ordinal Logistic Regression Models – CCR Models for multiple dichotomous/ordinal outcome variables under development

CCR-Cox: Survival Models – Extensions with Latent Class modeling being explored

Researchers interested in beta testing COREXpress™ should email:

will@statisticalinnovations.com

# 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.

Lyles R.H., Y. Guo and A. Hill (2009). "A Fresh Look at the Discrimination Function Approach for Estimating Crude or Adjusted Odds Ratios", The American Statistician, Vol 63, No. 4 (November ), pp 320-327.

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

Magidson, J. (1996). "Maximum Likelihood Assessment of Clinical Trials Based on an Ordered Categorical Response.", Drug Information Journal, Maple Glen, PA: Drug Information Association, Vol. 30, No. 1, 143-170.

Magidson, J., K. Wassmann, W. Oh, R. Ross, P. Kantoff, (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.