

Correlated Component Regression: Re-thinking Regression in the Presence of Near Collinearity

Jay Magidson

Abstract We introduce a new regression method—called Correlated Component Regression (CCR)—which provides reliable predictions even with near multicollinear data. Near multicollinearity occurs when a large number of correlated predictors and relatively small sample size exists as well as situations involving a relatively small number of correlated predictors. Different variants of CCR are tailored to different types of regression (e.g. linear, logistic, Cox regression). We also present a step-down variable selection algorithm for eliminating irrelevant predictors. Unlike PLS-R and penalized regression approaches, CCR is scale invariant. CCR is illustrated in several examples involving real data and its performance is compared with other approaches using simulated data.¹

Key words: Correlated component regression, Multicollinearity, High dimensional data, Big data, PLS regression, Variable selection, Suppressor variables, Scale invariance, Cross-validation

1 Background and Introduction

When correlation between predictor variables is moderate or high, coefficients estimated using traditional regression techniques become unstable or cannot be uniquely estimated due to multicollinearity (singularity of the covariance matrix). In the case of high dimensional data, where the number of predictor variables P approaches or exceeds the sample size N , such instability is often accompanied by perfect or near perfect predictions within the analysis sample. However, this seem-

¹ All data sets are available on the website statisticalinnovations.com

J. Magidson (✉)
Statistical Innovations Inc., Belmont, MA, USA
e-mail: jay@statisticalinnovations.com

ingly good predictive performance is usually associated with *overfitting*, and tends to deteriorate when applied to new cases outside the sample.

The primary “regularization” approaches that have been proposed for dealing with this problem are (1) penalized regression such as Ridge, Lasso and Elastic Net, and (2) dimension reduction methods such as Principle Component Regression, and PLS Regression (PLS-R). In this paper we describe a new method similar to PLS-R called Correlated Component Regression (CCR) and an associated step-down algorithm for reducing the number of predictors in the model to $P^* < P$. CCR has different variants depending upon the scale type of the dependent variable (e.g. CCR-linear regression for Y continuous, CCR-logistic regression for Y dichotomous, CCR-Cox regression for survival data). Unlike the other regularization approaches, the CCR algorithm shares with traditional maximum likelihood regression approaches the favorable property of scale invariance.

In this paper we introduce CCR, and describe its performance on various real and simulated data sets. The basic CCR algorithms are described in Sect. 2. CCR is contrasted with PLS-R in a linear regression key driver application with few predictors (Sect. 3) and in an application with Near Infrared (NIR) data involving many predictors (Sect. 4). We then describe the CCR extension to logistic regression, linear discriminant analysis (LDA) and survival analysis and discuss results from simulated data where suppressor variables are included among the predictors (Sect. 5). Results from our simulations suggest that CCR may be expected to outperform other sparse regularization approaches, especially when important suppressor variables are included among the predictors. We conclude with a discussion of a hybrid latent class CCR model extension (Sect. 6).

2 Correlated Component Regression

CCR utilizes $K < P$ correlated components, in place of the P predictors to predict an outcome variable. Each component S_k is an exact linear combination of the predictors, $X = (X_1, X_2, \dots, X_P)$, the first component S_1 capturing the effects of those predictors that have direct effects on the outcome. The CCR-linear regression (CCR-LM) algorithm proceeds as follows:

Estimate the *loading* $\lambda_g^{(1)}$, on S_1 , for each predictor $g = 1, 2, \dots, P$, as the simple regression coefficient in the regression of Y on X_g , $\lambda_g^{(1)} = \frac{\text{cov}(Y, X_g)}{\text{var}(X_g)}$. Then S_1 is defined as a weighted average of all 1-predictor effects:

$$S_1 = \frac{1}{P} \sum_{g=1}^P \lambda_g^{(1)} X_g \quad (1)$$

The predictions for Y in the 1-component CCR model are obtained from the simple OLS regression of Y on S_1 . Similarly, predictions for the 2-component CCR model are obtained from the simple OLS regression of Y on S_1 and S_2 , where the

second component S_2 , captures the effects of suppressor variables that improve prediction by removing extraneous variation from one or more predictors that have direct effects. Component $S_{k'}$ for $k' > 1$, is defined as a weighted average of all 1-predictor partial effects, where the partial effect for predictor g is computed as the partial regression coefficient in the OLS regression of Y on X_g and all previously computed components $S_k, k = 1, \dots, k' - 1$. For example, for $K = 2$ we have:

$$Y = \alpha + \gamma_{1,g}^{(2)} S_1 + \lambda_g^{(2)} X_g + \varepsilon_g^{(2)} \quad (2)$$

and $S_2 = \frac{1}{P} \sum_{g=1}^P \lambda_g^{(2)} X_g$, or more simply² we can write $S_2 = \sum_{g=1}^P \lambda_g^{(2)} X_g$.

As mentioned earlier, predictions for Y in the K -component CCR model are obtained from the OLS regression of Y on S_1, \dots, S_K . For example, for $K = 2$: $\hat{Y} = \alpha^{(2)} + b_1^{(2)} S_1 + b_2^{(2)} S_2$. In general, K^* components are computed, where the optimal value, K^* , is determined by M -fold cross-validation (CV). For $K = 1$, maximum regularization, no predictor correlation information is used in parameter estimation. As K is repeatedly incremented by 1, more and more information provided by the predictor correlations is utilized, and M -fold CV determines the value of K where near multicollinearity begins to deteriorate the predictive performance, the value for K^* being obtained accordingly. Deterioration occurs beginning at $K = 3$ for the example illustrated in Sect. 3, and thus $K^* = 2$.

Any K -component CCR model can be re-expressed to obtain regression coefficients for X by substituting for the components as follows:

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

Thus, the regression coefficient β_g for predictor X_g is simply the weighted sum of the loadings, where the weights are the regression coefficients for the components (component weights) in the K -component model: $\beta_g = \sum_{k=1}^K b_k^{(K)} \lambda_g^{(k)}$.

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 $\beta_g^* = (\sigma_g / \sigma_Y) \beta_g$, where σ denotes the standard deviation. M -fold CV is used to determine the two tuning parameters: the number of components K and number of predictors P .

Consider an example with 6 predictors. For any given value for K , and say $M = 10$ folds, the basic CCR algorithm is applied 10 times, generating predictions for cases in each of the 10 folds³ based on models with all 6 predictors, yielding a baseline (iteration = 0) $CV-R^2(K)$ for $P = 6$. In iteration 1, the variable reduction algorithm eliminates 1 predictor, which may not be the same predictor in all 10

² Going forward, the factor $1/P$ will be omitted which will not alter the predictions since multiplying S_k by P is offset by the OLS estimate for gamma (i.e., $\gamma_{k,g}^{(K)}$ becomes $\gamma_{k,g}^{(K)}/P$).

³ The square of the correlation between these predictions and the observed Y yields $CV-R^2$.

subsamples, each resulting 5-predictor model being used to obtain new predictions for the associated omitted folds, yielding $CV-R^2(K)$ for $P = 5$. In iteration 2, the variable reduction process continues resulting in 10 4-predictor models, which yields $CV-R^2(K)$ for $P = 4$. Following the last iteration, $P^*(K)$ is determined as the value of P associated with the maximum $CV-R^2(K)$.

The basic idea is that by applying the proper amount of regularization through the tuning of K , we reduce any confounding effects due to high predictor correlation, thus obtaining more interpretable regression coefficients, and better, more reliable predictions. In addition, tuning P tends to eliminate irrelevant or otherwise extraneous predictors and further improve both prediction and interpretability.

Since the optimal P may depend on K , P should be tuned for each K , the optimal (P^*, K^*) yielding the global maximum for $CV-R^2$. Alternatively, as a matter of preference a final model may be based on a smaller value for P and/or K , such that the resulting $CV-R^2$ is within c standard errors of the global maximum, where $c \leq 1$.

Since K can never exceed P , for $P = K$, the model becomes saturated and is equivalent to the traditional regression model.⁴ For pre-specified K , when P is reduced below K , we maintain the saturated model by also reducing K so $K = P$. For example, for $K = 4$, when we step down to 3 predictors, we reduce K so $K = 3$. Similarly, when we step down to 1 predictor, $K = 1$. This is similar to traditional stepwise regression with backwards elimination.

Prime predictors, those having direct effects, are identified as those having substantial loadings on S_1 , and suppressor variables, as those having substantial loadings on one or more other components, and relatively small loadings on S_1 . See Sect. 5 for further insight into suppressor variables.

Since CCR is scale invariant, it yields identical results regardless of whether predictions are based on unstandardized or standardized predictors (Z -scores). Other methods such as PLS-R and penalized regression (Ridge Regression, Lasso, Elastic Net) are not scale invariant and hence yield different results depending on the predictor scaling used.

3 A Simple Example with Six Correlated Predictors

Our first example makes use of data involving the prediction of car prices (Y) as a linear function of 6 predictors, each having a statistically significant positive correlation with Y (between 0.6 and 0.9).

- $N = 24$ car models
- Dependent variable: $Y = \text{PRICE}$ (car price measured in francs)
- 6 Predictor Variables:
 - $X_1 = \text{CYLINDER}$ (engine measured in cubic centimeters)
 - $X_2 = \text{POWER}$ (horsepower)
 - $X_3 = \text{SPEED}$ (top speed in kilometers/hour)

⁴ See Appendix for proof of this equivalence.

- X_4 = WEIGHT (kilograms)
- X_5 = LENGTH (centimeters)
- X_6 = WIDTH (centimeters)

The OLS regression solution (Table 1a) imposes no regularization, maximizing R^2 in the training sample. This solution is equivalent to that obtained from a saturated ($K = P = 6$ components) CCR model. Since this solution is based on a relatively small sample and correlated predictors, it is likely to overfit the data and the R^2 is likely to be an overly optimistic estimate of the true population R^2 . Table 1a shows only 1 statistically significant coefficient (0.05) and unrealistic (negative) coefficient estimates for 3 of the 6 predictors, which are problems that can be explained by model overfitting due to imposing no regularization.

Table 1: (a) (left) shows OLS Regression Coefficient results ($P = K = 6$) and (b) (right) shows R^2 and $CV\text{-}R^2$ for different numbers of components K and for the final CCR model ($P = 3, K = 2$)

	Unstandardized coefficients		Standardized	t	Sig.	P	K	R^2	$CV\text{-}R^2$
	$\hat{\beta}$	Std. error	$\hat{\beta}^*$						
CYLINDER	-1.9	33.6	-0.02	-0.06	0.95	6	1	0.7852	0.7457
POWER	1,315.9	613.5	0.89	2.14	0.05	6	2	0.8189	0.7461
SPEED	-472.5	740.3	-0.21	-0.64	0.53	6	3	0.8449	0.6732
WEIGHT	45.9	100.0	0.18	0.46	0.65	6	4	0.8469	0.6455
LENGTH	209.6	504.2	0.15	0.42	0.68	6	5	0.8474	0.6371
WIDTH	-505.4	1,501.6	-0.07	-0.34	0.74	6	6	0.8474	0.6342
(Constant)	12,070.4	194,786.6		0.06	95	3	2	0.8362	0.7690

To determine the value for K that provides the optimal amount of regularization, we choose the CCR model that maximizes the $CV\text{-}R^2$. For cross-validation we used 10 rounds of 6-folds, since 24 divides evenly into 6, each fold containing exactly 4 cars. Table 1b shows that $K = 2$ components provides the maximum $CV\text{-}R^2$ based on $P = 6$ predictors, and when the step-down algorithm is employed, $CV\text{-}R^2$ increases to 0.769 which occurs with $P^* = 3$ predictors.⁵ While traditional OLS regression yields a higher R^2 in the analysis sample (0.847 vs. 0.836), the 2-component CCR model with 3 predictors yields a higher $CV\text{-}R^2$, suggesting that this CCR model will outperform⁶ OLS regression when applied to new data.

Further evidence of improvement for the 2-component models over OLS regression is that the coefficients are more interpretable. Table 2 shows that the coefficients in the 2-component CCR models are all positive, which is what we would expect if we were to interpret them as measures of effect.⁷

⁵ The analysis was conducted using the CORExpress[®] package (patent pending) [1].

⁶ Since multiple rounds of 6-folds are performed, standard errors are available, which yield 95% confidence intervals for $CV\text{-}R^2$ of 0.746 ± 0.04 for the CCR model with 6 predictors and 0.769 ± 0.056 for the 3-predictor CCR model.

⁷ Interestingly, each CCR model based on an insufficient amount of regularization ($K > 2$) provides uninterpretable coefficients, in each case exactly three coefficients turning out negative.

Table 2 Comparison of results from PLSR (a) (left) with unstandardized predictors, and (b) with standardized predictors, and CCR (c) without variable selection and (d) (right) with variable selection

	PLS with unstandardized predictors ($K^* = 3$)	PLS with standardized predictors ($K^* = 2$)	CCR with $(K^* = 2)$	CCR with selection ($K^* = 2$)
Training R^2	0.83	0.81	0.82	0.84
CV- R^2	0.69	0.76	0.75	0.77
Predictors	$\hat{\beta}^*$	$\hat{\beta}^*$	$\hat{\beta}^*$	$\hat{\beta}^*$
CYLINDER	-0.02	0.19	0.19	0.00
POWER	0.43	0.31	0.37	0.45
SPEED	0.17	0.22	0.20	0.10
WEIGHT	0.48	0.18	0.17	0.44
LENGTH	-0.05	0.08	0.02	0.00
WIDTH	0.00	0.01	0.05	0.00

PLS-R with standardized predictors, the recommended PLS-R option when predictors are measured in different units, yields similar results to CCR here. When the predictors remain unstandardized, PLS-R yields more components ($K^* = 3$), two negative coefficients, and substantially worse predictions ($CV-R^2 = 0.69$), as the much larger variance for the predictor CYLINDER causes this predictor to dominate the first component, requiring two additional components to recover.

4 An Example with Near Infrared (NIR) Data

Next, we analyze high dimensional data involving $N = 72$ biscuits, each measured at each of $P = 700$ near infrared (NIR) wave-lengths corresponding to every other wavelength between the range 1,100–2,500 [2]. Since all 700 predictors are measured in comparable units in this popular PLS-R application, typically the 700 predictors are analyzed on an unstandardized basis, or standardized using Pareto scaling [3] where the scaling factor is the square root of the standard deviation. As shown above, results from PLS-R differ depending upon whether the predictors are standardized or not, while for the scale invariant CCR, no decision needs to be made regarding such standardization, predictions being identical in either case.

The goal of modeling here is to reduce costs of monitoring fat content by predicting the percent fat based on spectroscopic absorbance variables from the NIR frequencies. Following Kraemer and Boulesteix [4], we use $N = 40$ samples as the calibration (training) set to develop models based on the 700 wave lengths.

It is well known that for NIR data, a column plot of regression coefficients exhibit a sequence of oscillating patterns, the most important wavelength ranges being those with the highest peak-to-peak amplitude. For example, for these data, wavelengths in the 1,500–1,598 range yield a peak to peak amplitude of $0.109 - (-0.203) = 0.312$, based on a CCR model with $K = 9$ (see Fig. 1).

Table 3a compares the corresponding amplitudes obtained from CCR and both unstandardized and Pareto standardized PLS-R models, where the number of components is determined based on 10 rounds of 5-folds. As can be seen in Table 3a, all

three models agree that absorbances from the 1,500–1,598 wavelengths tend to be among the most important (relatively large amplitude).

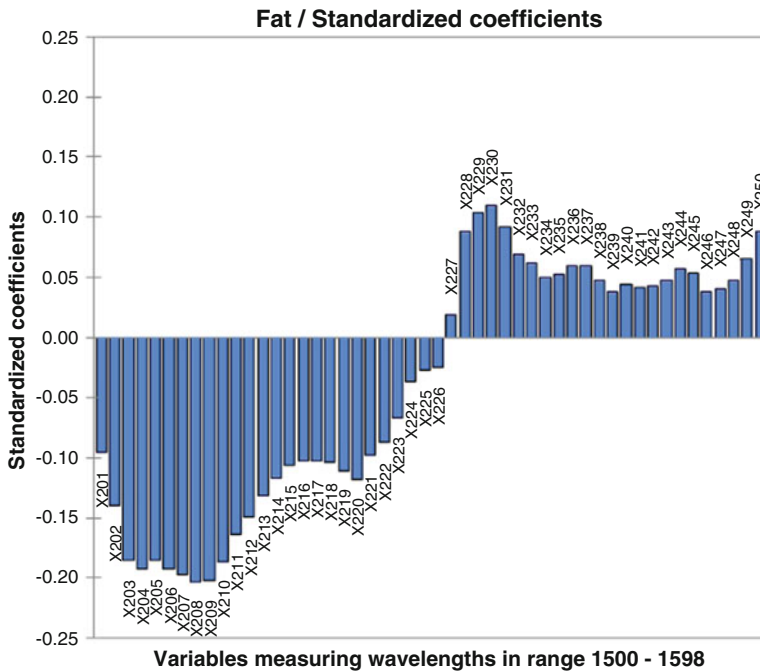


Fig. 1: Column plot of standardized coefficients output from XLSTAT-CCR

Previous analyses of these data excluded the highest 50 wavelengths since they were "... thought to contain little useful information" [5]. Table 3a shows that CCR identifies these wavelengths as least important (smallest amplitude), but the amplitude of 0.44 resulting from PLS-R suggests that these wavelengths are important.

Figure 2 shows the standardized coefficients for the 50 highest wavelengths for CCR and PLS-R models. As can be seen, the weights obtained from the CCR model are small and diminishing, the coefficients for the highest wavelengths being very close to 0. In contrast, PLS-R weights are quite high and show no sign of diminishing for the highest wavelengths (Fig. 2(right)), a similar pattern being observed for PLS-Pareto.

One possible reason that the conclusions from CCR and PLS-R differ regarding the importance of these high wavelengths is that its scale invariance property allows CCR to better determine that the high variability associated with these wavelengths is due to increased amounts of measurement error. In other words, the much higher amplitude obtained from PLS-R is likely due to the higher standard deviations of the absorbances in this range.

Table 3: (a) (left) Comparison of peak-to-peak amplitudes for various frequency ranges based on three models, with the most and least important ranges according to CCR in bold, and (b) (right) comparison of CV- R^2 (highest is bold) obtained from three models with ($P = 700$) and without ($P = 650$) the highest wavelengths included among the predictors

Wavelengths	Peak-to-peak amplitude based on standardized coefficients			CCR		PLSR		PLS-Pareto		
	CCR	PLSR	PLS-Pareto	K	$P = 700$	$P = 650$	$P = 700$	$P = 650$	$P = 700$	$P = 650$
	($K = 9$)	($K = 13$)	($K = 13$)							
1,100–1,198	0.16	0.12	0.19	1	0.237	0.232	0.260	0.257	0.247	0.245
1,200–1,298	0.24	0.15	0.24	2	0.506	0.589	0.345	0.461	0.412	0.477
1,300–1,398	0.11	0.08	0.13	3	0.759	0.860	0.736	0.725	0.721	0.736
1,400–1,498	0.27	0.25	0.21	4	0.914	0.932	0.906	0.835	0.922	0.882
1,500–1,598	0.31	0.31	0.32	5	0.948	0.946	0.916	0.928	0.933	0.917
1,600–1,698	0.23	0.14	0.15	6	0.948	0.951	0.919	0.947	0.927	0.949
1,700–1,798	0.27	0.24	0.22	7	0.945	0.947	0.930	0.942	0.936	0.946
1,800–1,898	0.20	0.15	0.17	8	0.955	0.953	0.936	0.938	0.944	0.948
1,900–1,998	0.07	0.47	0.36	9	0.962	0.960	0.932	0.952	0.946	0.952
2,000–2,098	0.22	0.37	0.30	10	0.960	0.963	0.939	0.958	0.946	0.961
2,100–2,198	0.16	0.17	0.15	11	0.957	0.959	0.942	0.959	0.951	0.962
2,200–2,298	0.18	0.30	0.29	12	0.958	0.959	0.949	0.958	0.952	0.961
2,300–2,398	0.18	0.55	0.47	13	0.958	0.959	0.950	0.956	0.954	0.9059
2,400–2,498	0.06	0.44	0.25	14	0.958	0.958	0.947	0.953	0.953	0.957
				15	0.958	0.957	0.946	0.952	0.952	0.956

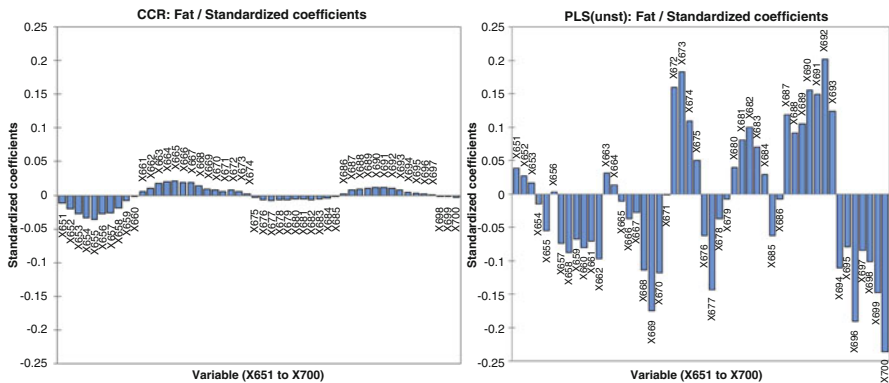


Fig. 2: Comparison of column plots of standardized coefficients for 50 highest wavelengths based on the CCR (left) vs. PLS-R estimated with unstandardized predictors (right)

To test the hypothesis that these higher wavelengths tend to be unimportant, we re-estimated the models after omitting these variables. Table 3b shows that for all three models, the $CV-R^2$ increases when these variables are omitted, supporting the hypothesis that these wavelengths are not important.

In order to compare the predictive performance of CCR with other regularization approaches, 100 simulated samples of size $N = 50$ were generated with 14 predictors according to the assumptions of OLS regression. An additional 14 extraneous predictors, correlated with the 14 true predictors, plus 28 irrelevant predictors, were also generated and included among the candidate predictors. The results indicated that CCR outperformed PLS-R, Elastic Net, and sparse PLS with respect to mean squared error, and several other criteria. All methods were tuned using an independent validation sample of size 50 (for more details, see [6]).

5 Extension of CCR to Logistic Regression, Linear Discriminant Analysis and Survival Analysis

When the dependent variable is dichotomous, the CCR algorithm generalizes directly to CCR-LOGISTIC and CCR-LDA respectively depending upon whether no assumptions are made about the predictor distributions, or whether the normality assumptions from linear discriminant analysis are made. In either case, the generalization involves replacing Y by $Logit(Y)$ on the left side of the linear equations. Thus, for example, under CCR-LOGISTIC and CCR-LDA Eq. 2 becomes:

$$Logit(Y) = \alpha + \gamma_{1,g}^{(2)} S_1 + \lambda_g^{(2)} X_g \quad (3)$$

where parameter estimation in each regression equation is performed by use of the appropriate ML algorithm (for logistic regression or LDA).

M -fold cross-validation continues to be used for tuning, but $CV-R^2$ is replaced by the more appropriate statistics CV -Accuracy and CV -AUC, AUC denoting the Area Under the ROC Curve. Accuracy is most useful when the distribution of the dichotomous Y is approximately uniform, about 50% of the sample being in each group. When Y is skewed, accuracy frequently results in many ties and thus is not as useful. In such cases AUC can be used as a tie breaker with Accuracy as the primary criterion or in the case of large skew, AUC can replace accuracy as primary.

For survival data, Cox regression and other important log-linear hazard models can be expressed as Poisson regression models since the likelihood functions are equivalent [7]. As such, CCR can be employed using the logit equation above where Y is a dichotomous variable indicating the occurrence of a rare event. In this case since Y has an extreme skew, the AUC is used as the primary criterion.

Similar to the result for CCR-linear regression, predictions obtained for the saturated CCR model for dichotomous Y are equivalent to those from the corresponding traditional model (logistic regression, LDA and Poisson regression).⁸ In addition, for

⁸ In general, the saturated model occurs when $K \geq \text{minimum}(P, N - 1)$.

dichotomous Y the 1-component CCR model is equivalent to Naïve Bayes, which is also called diagonal discriminant analysis [8] in the case of CCR-LDA.

In a surprising result reported in [9], for high dimensional data (small samples and many predictors) generated according to the LDA assumptions, traditional LDA does not work well, and is outperformed by Naïve Bayes. Because of the equivalences described above, this means that the *1-component* CCR model should outperform the *saturated* CCR model under such conditions. However, we know that the Naïve Bayes model will not work well if predictors include 1 or more important suppressor variables, since suppressor variables tend to have 0 loadings on the first component and require at least two components for their effects to be captured in the model [10]. Thus, a CCR model with two components should outperform Naïve Bayes whenever important suppressor variables are included among the predictors.

Despite extensive literature documenting the enhancement effects of suppressor variables (e.g. [11, 12]), most pre-screening methods omit suppressor variables prior to model development, resulting in suboptimal models.⁹ Since suppressor variables are commonplace and often are among the most important predictors in a model [10], such screening is akin to “throwing out the baby with the bath water.”

In order to compare the predictive performance of CCR with other sparse modeling methods in a realistic high dimensional setting, data were simulated according to LDA assumptions to reflect the relationships among real world data for prostate cancer patients and normals where at least one important suppressor variable was among the predictors. The simulated data involved 100 samples each with $N=25$ cases in each group, the predictors including 28 valid predictors plus 56 that were irrelevant. The sparse methods included CCR, sparse PLS-R [13, 14] and the penalized regression methods Lasso and Elastic Net [15–17]. For tuning purposes, cross-validation with five folds was used with accuracy as the criterion for all methods.

Results showed that CCR with typically 4–10 components outperformed the other methods with respect to accuracy (82.6% vs. 80.9% for sparse PLS-R, and under 80% for Lasso and Elastic Net), and fewest irrelevant predictors (3.4 vs. 6.2 for Lasso, 11.5 for Elastic Net and 13.1 for sparse PLS-R). The most important variable, which was a suppressor variable, was captured in the CCR model in 91 of the 100 samples compared to 78 for sparse PLS-R, 61 for elastic net and only 51 for Lasso. For further details of this and other simulations see [6].

6 Extension to Latent Class Models

In practice, sample data often reflects two or more distinct subpopulations (latent segments), with different intercepts and/or different regression coefficients, possibly due to different key drivers or at least different effects for the key drivers. In this section we describe a 2-step hybrid approach for identifying the latent segments without use of the predictors (step 1) and then using CCR to develop a predictive

⁹ For a rare exception, ISIS (see [19]) corrects for the exclusion of suppressor variables by the popular SIS screening. CCR has been shown to outperform ISIS in a simulation study [10].

model based on a possibly large number of predictors (step 2). If the predictors are characteristics of the respondents, then the dependent variable (Y) would be the latent classes, while if the predictors were attributes of objects being rated, Y would be taken as the ratings.

As an example of the first case where the latent segments have different intercepts, in step 1 a latent class (LC) survival analysis was conducted on a sample of patients with late stage prostate cancer. The LC model identified both long-term and short term survival groups [18]. The goal in that study was to use gene expression measurements to predict whether patients belong to the longer or shorter survival class. Since the relevant genes were not known beforehand, the large number of available candidate predictors (genes) ruled out use of traditional methods.

In this case, CCR can be used to simultaneously select the appropriate genes and develop reliable predictions of LC membership based on the selected genes. One way to perform this task is to predict the dichotomy formed by the two groups of patients classified according to the LC model. However, this approach is suboptimal because the classifications contain error due to modal assignment. That is, assigning patients with a posterior probability of say 0.6 of being a long term survivor to this class (with probability 1) ignores the 40% expected misclassification error ($1 - 0.6 = 0.4$). The better way is to perform a weighted logistic (or LDA) CCR regression, where posterior probabilities from the LC model serve as case weights.

Table 4: Results from CCR showing that $P = 3$ of the 16 attributes were selected for inclusion in the model together with the random intercept CFactor1

Results for segment 1		Results for segment 2	
Variable	Standardized coefficient	Variable	Standardized coefficient
CFactor1	0.425	CFactor1	0.555
Fructose	-0.128	Sweeteningpower	-0.169
Sweeteningpower	0.238	Smellintensity	-0.129
Acidity	-0.325	Acidity	0.214

As an example of the second case, consider ratings on 6 different orange juice (OJ) drinks provided by 96 judges [20]. Based on these ratings, in step 1 a LC regression determines that there are two latent segments¹⁰ exhibiting different OJ preferences. In step 2, separate weighted least squares CCR regressions are performed for each class to predict ratings based on the 16 OJ attributes. For a given class, posterior membership probabilities for that class are used as case weights.

For this application CCR is needed because traditional regression can include no more than six attributes in the model due to the fact that the attributes describe the six juices rather than the respondents. In addition, since these data consist of multiple records (6) per case, residuals from records associated with the same case are correlated, a violation of the independent observations assumption. This violation

¹⁰ The number of classes was determined based on the Bayesian Information Criterion. For further details of this methodology, see [21].

is handled in step 1 by the LC model satisfying the “local independence” assumption. In step 2, the cross-validation is refined by assigning records associated with the same case to the same fold. Separate CCR models are developed for each LC segment, and then combined to obtain predicted ratings, providing substantial improvement over the traditional regression ($CV-R^2$ increases from 0.28 to 0.48). Results of step 2 are summarized in Table 4, showing that the most important attribute for both segments is acidity since it has the highest standardized coefficient magnitude. Segment 1 tends to prefer juices with low acidity (negative coefficient) and high sweetening power (positive coefficient) while the reverse is true for segment 2. Details of this analysis are provided in tutorials from www.statisticalinnovations.com.

Acknowledgements The author gratefully acknowledges the contributions of Yiping Yuan in conducting the simulations and providing the proof in the Appendix, Michel Tennenhaus for contributing the car price and orange juice example datasets, programming assistance and ongoing support from Jeremy Magland, Alexander Ahlstrom, William Barker, and helpful comments from Harald Martens, Vincenzo Vinzi, Anthony Babinec, Gary Bennett, Karl Wassmann, David Rindskoff, Nicole Krämer, and an anonymous reviewer.

Appendix

Claim: OLS predictions based on X are equivalent to predictions based on $S = XA$, where A is a nonsingular matrix.

Proof:

- Predictions based on X :

$$\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y.$$

- Predictions based on S :

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

Equations 4 and 5 above follow from standard operations with square matrices:

$$(BC)' = C'B' \quad \text{and} \quad (BC)^{-1} = C^{-1}B^{-1}.$$

It also follows that the OLS regression coefficients for X are identical to those obtained from CCR with a saturated model (i.e., $K = P$).

References

- [1] J. Magidson, "CORExpress User's Guide: Manual for CORExpress", Belmont, MA: Statistical Innovations Inc., 2011.
- [2] B. Osbourne, T. Fearn, A. Miller, and S. Douglas, "Application of near infrared reflectance spectroscopy to compositional analysis of biscuits and biscuit dough," *Journal of Science and Food Agriculture*, 35, 99–105, 1984.
- [3] L. Eriksson, E. Johansson, N. Kettaneh-Wold, and S. Wold, "Introduction to multi- and megavariate data analysis using projection methods (PCA & PLS)," *Umetrics*, pp. 213–225, 1999.
- [4] N. Kraemer, and A. Boulesteix, "Penalized Partial Least Squares (PPLS)," R Package, V. 1.05, Aug. 2011.
- [5] P.J. Brown, T. Fearn, and M. Vannucci, "Bayesian wavelet regression on curves with application to a spectroscopic calibration problem," *Journal of the American Statistical Association*, 96, 398–408, 2001.
- [6] J. Magidson, "Correlated Component Regression: A Prediction/Classification Methodology for Possibly Many Features," *JSM Proceedings of the American Statistical Association*, pp. 4372–4386, 2010.
- [7] N. Laird and D. Oliver, "Covariance analysis of censored survival data using log-linear analysis techniques," *Journal of the American Statistical Association*, 76, pp. 231–240, 1981.
- [8] T.R. Golub, D.K.Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, 286, pp. 531–537, Oct. 1999.
- [9] P. Bickel and E. Levina, "Some theory for Fisher's linear discriminant function, 'naïve Bayes' and some alternatives when there are many more variables than observations," *Bernoulli*, 10, 989–1010, 2004.
- [10] J. Magidson and K. Wassmann, "The Role of Proxy Genes in Predictive Models: An Application to Early Detection of Prostate Cancer," *2010 JSM Proceedings of American Statistical Association, Biometrics Section*, pp. 2739–2753, 2010.
- [11] P. Horst, "The role of predictor variables which are independent of the criterion," *Social Science Research Bulletin*, 48, pp. 431–436, 1941.
- [12] H. Lynn, "Suppression and Confounding in Action," *The American Statistician*, 57, pp. 58–61, 2003.
- [13] H. Chun and S. Keleş, "Sparse partial least squares regression for simultaneous dimension reduction and variable selection," University of Wisconsin, Madison, 2009.
- [14] H. Chun and S. Keleş, "Sparse Partial Least Squares Classification for High Dimensional Data," *Statistical Applications in Genetics and Molecular Biology*, 9, 17, 2010.
- [15] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society Series B (Methodological)*, 58, pp. 267–288, 1996.
- [16] J. Friedman, T.Hastie, and R.Tibshirani, "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software*, 33, pp. 1–22, 2010.
- [17] J. Friedman, T. Hastie, and R. Tibshirani, "Lasso and elastic-net regularized generalized linear models," Version 1.3, Jstatsoft.org, April 25, 2010.
- [18] R. Ross, M. Galsky, H. Scher, J. Magidson, K. Wassmann, G. Lee, L. Katz, S. Subudhi, A. Anand, M. Fleisher, P. Kantoff, and W. Oh, "A whole-blood RNA transcript-based prognostic model in men with castration-resistant prostate cancer: a prospective study," *Lancet Oncology*, 2012; [http://dx.doi.org/10.1016/S1470-2045\(12\)70263-2](http://dx.doi.org/10.1016/S1470-2045(12)70263-2).

- [19] J. Fan, Samworth, and W. Yichao, "Ultrahigh Dimensional Feature Selection: Beyond the Linear Model," *Journal of Machine Learning Research*, 10, pp. 2013–2038, 2009.
- [20] M Tenenhaus, M., Pagès, J., Ambroisine L. and C. Guinot, "PLS methodology for studying relationships between hedonic judgments and product characteristics," *Food Quality and Preference*, 16, pp. 315–325, 2005.
- [21] J. Magidson, and J. Vermunt, "Latent Class Models," in D.Kaplan (Ed.), *The Sage Handbook of Quantitative Methodology for the Social Sciences*, pp. 175–198. Thousand Oaks: Sage Publications, 2004.