Reprinted from 2013 Springer Verlag book: "New perspectives in Partial Least Squares and Related Methods"

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

Jay Magidson
Statistical Innovations Inc., e-mail: jay@statisticalinnovations.com

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

The principle "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, the first component S_1 capturing the effects of 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 $g = 1, 2, ..., P$, as the simple regression coefficient in the regression of Y on X_g , $\lambda_g^{(1)} = \frac{cov(Y,X_G)}{var(X_g)}$. Then S_1 is defined as a weighted average of all 1-predictor effects:

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

The predictions for Y in the 1-component CCR model are obtained from the simple OLS regression of Y on S_1. Similarly, the 2-component CCR model is formed by the simple OLS regression of Y on S_1 and S_2, where 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. $S_{k'}$ 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 $S_k, k = 1, ..., k' - 1$. For example, for $K = 2$ we have:

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

and $S_2 = \frac{1}{P} \sum_{g=1}^{P} \lambda_g^{(2)} X_g$

The predictions for Y in the K-component CCR model are obtained from the OLS regression of Y on $S_1, ..., S_K$. For example, for $K = 2$: $\hat{Y} = \alpha^{(2)} + b_1^{(2)} S_1 + b_2^{(2)} S_2$ Similarly, additional components are computed as required, and as illustrated in Sect. 3, M-fold cross-validation (CV) can be used to determine the optimal number of components K*.

Any K-component CCR model can be re-expressed to obtain regression coefficients for the predictors in the final K-component model 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 2 tuning parameters – the number of components K and number of predictors P.

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

Since K can never exceed P, for $P = K$, the model becomes saturated and is equivalent to the traditional regression model.[2] For pre-specified K, to reduce P 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 predictor*s, those having direct effects, are identified as those having substantial loadings on S_1, and suppressor variables, as those having substantial loadings on S_2, and relatively small loadings on S_1. See Sect. 5 for further insight into the suppressor variables.

[2] See Appendix for proof.

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 scale of the predictors.

3 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 (between .6 and .9) with Y.

- N = 24 car models
- Dependent variable: Y = PRICE (car price measured in francs)
- 6 Predictor Variables:
  - X1 = CYLINDER (engine measured in cubic centimeters)
  - X2 = POWER (horsepower)
  - X3 = SPEED (top speed in kilometers/hour)
  - X4 = WEIGHT (kilograms)
  - X5 = LENGTH (centimeters)
  - X6 = WIDTH (centimeters)

The Ordinary Least Squares (OLS) regression solution (Table 1a) imposes no regularization, maximizing R^2 in the training sample. The OLS 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. Consistent with an overfit model, Table 1a shows only 1 significant coefficient and unrealistic (negative) coefficient estimates for 3 of the 6 predictors.

	Unstandardized Coefficients		Standardized Coefficients		
	$\hat{\beta}$	Std. Error	$\hat{\beta}$	t	Sig.
CYLINDER	**-1.9**	33.6	**-.02**	-.06	.95
POWER	1315.9	613.5	.89	2.14	.05
SPEED	**-472.5**	740.3	**-.21**	-.64	.53
WEIGHT	45.9	100.0	.18	.46	.65
LENGTH	209.6	504.2	.15	.42	.68
WIDTH	**-505.4**	1501.6	**-.07**	-.34	.74
(Constant)	12070.4	194786.6		.06	95

P	K	R^2	CV-R^2
6	1	0.7852	0.7457
6	2	0.8189	**0.7461**
6	3	0.8449	0.6732
6	4	0.8469	0.6455
6	5	0.8474	0.6371
6	6	0.8474	0.6342
3	2	0.8362	**0.7690**

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

To determine the value for K that provides the optimal amount of regularization, we choose the CCR model that maximizes the cross-validated 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-R^2 based on 6-predictor models, and when the step-down algorithm is employed, CV-R^2 increases to .769 which occurs with $P^* = 3$ predictors[3]. While traditional OLS regression yields a higher R^2 in the analysis sample (.847 vs. .836), the 2-component CCR model with 3 predictors yields a higher CV-R^2, suggesting that this CCR model will outperform the OLS regression model 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. [4]

Table 2 Coefficient estimates obtained from 2-component CCR models a. (left) without variable selection and b. (right) with variable selection.

Predictor	$\hat{\beta}$	$\hat{\beta}^*$
CYLINDER	20.9	0.19
POWER	545.5	0.37
SPEED	445.7	0.20
WEIGHT	43.4	0.17
LENGTH	32.6	0.02
WIDTH	343.6	0.05
(Constant)	-177941	

Predictor	$\hat{\beta}$	$\hat{\beta}^*$
POWER	673.3	0.45
SPEED	222.9	0.10
WEIGHT	110.9	0.44
(Constant)	-115044	

PLS-R yields similar results to CCR for these data when the predictors are standardized, the common PLS-R option undertaken when predictors are measured in different units. When the predictors remain unstandardized, PLS-R yields substantially worse results as the much larger variance for the predictor CYLINDER causes this predictor to dominate the 1st component.

4 Example with Near Infrared (NIR) data

Next, we analyze high dimensional data involving N = 72 cookies, each measured at each of P = 700 near infrared (NIR) wave-lengths corresponding to every other wavelength between the range 1100–2500 [3]. 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 [4] where the scaling factor is the square root of the standard deviation. As mentioned 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, the results being identical in either case.

[3] The analysis was conducted using the COREXpress® package (patent pending). [2]

[4] It is also interesting to note that each CCR model with less than the optimal amount of regularization (i.e., models for which $K > 2$) provides uninterpretable coefficients, in each case exactly 3 coefficients turning out negative.

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 1500-1598 ranges yield a peak to peak amplitude of $.109-(-.203)=.312$, based on a CCR model with $K=9$ (see Figure 1).

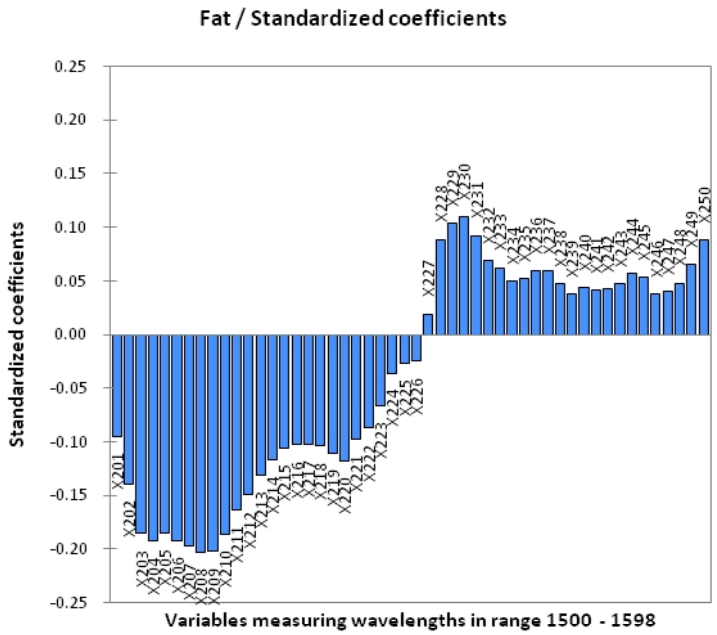

Fig. 1 Column plot of standardized coefficients output from XLSTAT-CCR program.

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. The Pareto standardization, where the absorbance variables are divided by the square root of their standard deviation, is generally recommended for NIR data [1]. As can be seen all 3 models agree that absorbances from the 1500-1598 wavelengths tend to be among the most important (relatively large amplitude).

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

	Peak-to-peak Amplitude based on Standardized Coefficients		
Wavelengths	CCR(K=9)	PLS-R (K=13)	PLS-Pareto(K=13)
1100-1198	0.16	0.12	0.19
1200-1298	0.24	0.15	0.24
1300-1398	0.11	0.08	0.13
1400-1498	0.27	0.25	0.21
1500-1598	**0.31**	**0.31**	**0.32**
1600-1698	0.23	0.14	0.15
1700-1798	0.27	0.24	0.22
1800-1898	0.20	0.15	0.17
1900-1998	0.07	0.47	0.36
2000-2098	0.22	0.37	0.30
2100-2198	0.16	0.17	0.15
2200-2298	0.18	0.30	0.29
2300-2398	0.18	0.55	0.47
2400-2498	**0.06**	**0.44**	0.25

	CCR		PLS-R		PLS-Pareto	
K	P=700	P=650	P=700	P=650	P=700	P=650
1	0.237	0.232	0.260	0.257	0.247	0.245
2	0.506	0.589	0.345	0.461	0.412	0.477
3	0.759	0.860	0.736	0.725	0.721	0.736
4	0.914	0.932	0.906	0.835	0.922	0.882
5	0.948	0.946	0.916	0.928	0.933	0.917
6	0.948	0.951	0.919	0.947	0.927	0.949
7	0.945	0.947	0.930	0.942	0.936	0.946
8	0.955	0.953	0.936	0.938	0.944	0.948
9	**0.962**	0.960	0.932	0.952	0.946	0.952
10	0.960	**0.963**	0.939	0.958	0.946	0.961
11	0.957	0.959	0.942	**0.959**	0.951	**0.962**
12	0.958	0.959	0.949	0.958	0.952	0.961
13	0.958	0.959	**0.950**	0.956	**0.954**	0.959
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

Table 3 a.(left) comparison of peak-to-peak amplitudes for various frequency ranges based on 3 models with the most and least important ranges according to CCR bolded, and b.(right) comparision of CV-R^2 (highest is bolded) obtaained from 3 models with (P=700) and without (P=650) the highest wavelength included among the predictors.

identifies these wavelengths as least important (smallest amplitude), but the amplitude of .44 resulting from PLS-R suggests that these wave-lengths 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 (Figure 2(right)), this pattern being similar for PLS-Pareto.

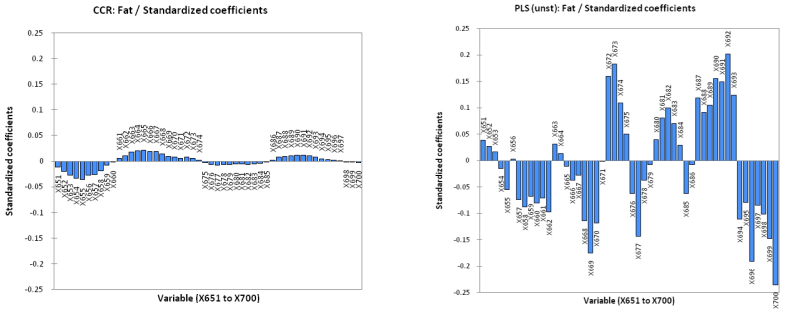

Fig. 2 Comparison of column plots of standardized coefficients for 50 highest wavelengths based on the CCR (left) vs. PLS-R (right).

One possible reason for the different results from CCR and PLS-R, is that due to its scale invariance property, CCR is better able than PLS-R to recognize that the higher variability associated with these high wavelengths is due to increased amounts of measurement error.[5]

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

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:

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

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 dependent variable Y is approximately uniform, about 50% of the sample being in each group. When Y is skewed, accuracy 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, an approximation of Cox regression can be provided by Poisson regression where survival is treated as a rare event [17]. 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). [6] In addition, for dichotomous Y the 1-component CCR model is equivalent to Naïve Bayes, which is also called diagonal discriminant analysis [7] in the case of CCR-LDA.

In a surprising result reported in [6], for high dimensional data (small samples and many predictors) generated according to the LDA assumptions, traditional LDA

[5] The higher amplitude obtained from PLS-R in this range is likely due to the fact that the standard deviation of the absorbances in this range are substantially higher than those in the other ranges.

[6] The saturated model occurs when $K > minimum(P, N-1)$.

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 1st component and require at least 2 components for their effects to be captured in the model [9]. Thus, a CCR model with 2 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.,* [14, 15]), most pre-screening methods omit suppressor variables prior to model development, resulting in suboptimal models. [7] Since suppressor variables are commonplace and often are among the most important predictors in a model [9], 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 [11] and the penalized regression methods Lasso and Elastic Net [12, 13]. For tuning purposes, cross-validation with 5 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 [10].

6 Extension to Latent Class Model

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 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.

[7] For a rare exception, ISIS (see [16]) corrects for the exclusion of suppressor variables by the popular SIS screening.

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 [8]. 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 simply predict the dichotomy formed by the 2 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 .6 of being a long term survivor to this class (with probability 1) ignores the 40% expected misclassification error ($1 - .6 = .4$). The better way is to perform a weighted logistic (or LDA) CCR regression, where the posterior probabilities from the LC model serve as case weights.

As an example of the second case, consider ratings on 6 different orange juice (OJ) drinks provided by each of 96 judges [18]. Based on these ratings, in step 1 a LC regression analysis determines that there are 2 latent segments 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, the 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 6 attributes in the model due to the fact that the attributes describe the 6 juices rather than the respondents. In addition, since these data consist of multiple (6) records per case, residuals from records associated with the same case are correlated, a violation of the independent observations assumption. This violation is handled in step 1 by the LC model satisfying the 'local independence' assumption. In step 2, the cross-validation is accomplished 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 (cross-validated R-square increases from .28 to .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. More complete details of this analysis are provided in tutorials on the website www.statisticalinnovations.com.

Results for Segment 1	
Variable	Standardized Coefficient
CFactor1	0.425
Fructose	-0.128
Sweeteningpower	0.238
Acidity	-0.325

Results for Segment 2	
Variable	Standardized Coefficient
CFactor1	0.555
Sweeteningpower	-0.169
Smellintensity	-0.129
Acidity	0.214

Table 4 Results from CCR showing that $P = 3$ of the 15 attributes were selected for inclusion in the model together with the random intercept CFactor1.

Appendix

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

Proof:

- Predictions base on X:

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

- Predictions base on S:

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

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

$$\begin{aligned}(BC)' &= C'B' \\ (BC)^{-1} &= C^{-1}B^{-1}\end{aligned}$$

References

1. 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.
2. J. Magidson, "COREXpress Users Guide: Manual for COREXpress", Belmont, MA: Statistical Innovations Inc., 2011.
3. 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.
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(454):398–408, 2001.
6. 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(6), 989–1010, 2004.
7. 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(5439): pp. 531–537, Oct. 1999.
8. R. Ross, M. Galsky, H. Scher, J. Magidson, K. Wassmann, G. Lee, L. Katz, S. Subudhi, A. Anand, M. Fleisher, P. Kantoff, W. Oh, "A whole-blood RNA transcript-based prognostic model in men with castration-resistant prostate cancer: a prospective study," *Lancet Oncology*, forthcoming, 2012.
9. 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.
10. J. Magidson, "Correlated Component Regression: A Prediction/Classification Methodology for Possibly Many Features," *2010 JSM Proceedings of American Statistical Association, Biometrics Section*, 2010.
11. H. Chun. and S. Keleş, "Sparse partial least squares regression for simultaneous dimension reduction and variable selection," University of Wisconsin, Madison, 2009.
12. J. Friedman, T.Hastie, R.Tibshirani, "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software*, 33(1), pp. 1–22, 2010.
13. J. Friedman, T. Hastie, and R. Tibshirani, "Lasso and elastic-net regularized generalized linear models," Version 1.3, Jstatsoft.org, April 25, 2010.
14. P. Horst, "The role of predictor variables which are independent of the criterion,"*Social Science Research Bulletin*, 48, pp. 431–436, 1941.
15. H. Lynn, "Suppression and Confounding in Action," *The American Statistician*, Vol.57, pp. 58–61, 2003.
16. J. Fan, Samworth, and W. Yichao, "Ultrahigh Dimensional Feature Selection: Beyond the Linear Model," *Journal of Machine Learning Research*, 10, pp. 2013–2038, 2009.
17. N. Laird, D. Oliver, "Covariance analysis of censored survival data using log-linear analysis techniques," *Journal of the American Statistical Association*, 76: pp. 231–240, 1981.
18. M Tenenhaus, M., Pags, J., Ambroisine L. and C. Guinot, "PLS methodology for studying relationships between hedonic judgments and product characteristics," *Food Quality and Preference*, 16, 4, pp. 315–325, 2005.