

CORRELATED COMPONENT REGRESSION (CCR) - A BRIEF METHODOLOGICAL DESCRIPTION

Jay Magidson and Gary Bennett – December 2011

Overview

At the heart of CCR is a simple but unique regularisation algorithm which is designed to produce more stable estimates of model parameters which cross-validate better than the un-stabilised estimates obtained from conventional models. This algorithm performs a special type of variable reduction of the most relevant predictor components to produce these stabilised estimates. Using cross-validation to determine which model to select ensures that the model will predict well on new cases rather than over-fitting the sample being used to develop the model. The algorithm uses a very efficient cross-validation (CV) procedure where cases in the sample are used sometimes for training and other times for validation purposes over a large number of iterations. Variable selection is based on a stepping-down procedure which initialises with the full model with all variables included and then eliminates variables with the smallest standardised coefficients one at a time, re-estimating the model at each step. By default, elimination of predictors continues until the CV performance criteria begins to decline

In essence the same approach is used for all of the types of predictive models currently implemented in CORExpress:

- linear regression
- logistic regression
- linear discriminant analysis (with 2 groups)
- survival models (Cox regression)

The main difference for each of these is:

1. The “work-horse” model form used to calculate model parameters at each stage (see any reference on standard statistical text for details of how these models are calculated).
2. The criteria used to evaluate how well the model predicts the training and cross-validation samples

Each of the three main components of the algorithm is described in detail in the remainder of this note.

Regularisation Algorithm

Summary

The Correlated Component Regression algorithm utilises K correlated components, where K is specified by the user or automatically determined by the program for a user-specified maximum number of predictors. As with Principal Components Analysis each component is an exact linear combination of some underlying vector of variables, in this case a vector of p predictor $X =$

(x_1, x_2, \dots, x_p) of a single dependent variable y . Unlike Principal Components where the weights for the components are chosen in order to explain the Correlation Matrix of X , the weights in CCR are chosen to maximise the components ability to predict y . We represent the K components as a vector $S = (S_1, S_2, \dots, S_K)$. Typically K (number of components) is less than P (number of predictors).

These components S can be thought of as composite predictors (derived from X) and are used as the terms in the “regularised” model in place of X in the predictive model. Component level coefficients are estimated. These can be further decomposed into regularised coefficients for X by using the fact that the components are exact linear combinations of X . We are therefore left with a regularised predictive equation of y on X . Optionally the components themselves can be interpreted in order to better understand the dimensions of the predictive equation.

The more regularisation required to obtain a stable model, the fewer the number of components in the final model with K approaching 1. In the case of logistic regression and linear regression, the CCR model with $K=1$ is equivalent to the special case known as Naive Bayes. For data requiring less regularisation more factors are justified with K approaching P . For the special case where $K=P$ the estimated model is equivalent to standard regression (linear or logistic) with no regularisation.

A detailed description of the estimation steps for the $1, 2, \dots, K$ component model are given below.

Step 1 – Estimation of first component

The first component is a weighted average of all P possible 1-predictor models of y on X . We estimate P one predictor models:

$$y = \alpha_g^{(1)} + \lambda_g^{(1)} x_g + \varepsilon_g^{(1)} \quad g = 1, 2, \dots, P$$

We then form the first component S_1 as a linear combination of X using the parameters λ_g from the 1-predictor models as weights (ignoring the intercepts α_g for now):

$$S_1 = \sum_{g=1}^P \lambda_g^{(1)} x_g$$

This component captures the direct effects of X for the regression y on X as a simple average of the one predictor models. We can estimate a one-component CCR model which is a regression of y on S_1 :

$$\hat{y} = \alpha^{(1)} + b_1^{(1)} S_1$$

where S_1 acts as a composite of the vector of predictors X . Note that (a) we have re-estimated a new intercept and (b) the parameter $b_1^{(1)}$ is the effect of the first component S_1 on y . We can decompose this into regularised effects for X on y using the fact that S_1 is a linear combination of X :

$$\hat{y} = \alpha^{(1)} + b_1^{(1)} \sum_{g=1}^P \lambda_g^{(1)} x_g$$

We can simplify this as:

$$\hat{y} = \alpha^{(1)} + \sum_{g=1}^p \lambda_g^{*(1)} x_g$$

Where $\lambda_g^{*(1)} = b_1^{(1)} \lambda_g^{(1)}$ is the regularised coefficient for x_g in the one component model.

Step 2 – Estimation of second component

The second component is defined as a weighted average of $\lambda_g^{(2)} x_g$ $g = 1, 2, \dots, P$

Where each $\lambda_g^{(2)}$ is estimated from the following 2-predictor model:

$$y = \alpha_g^{(2)} + \gamma_{1,g}^{(2)} S_1 + \lambda_g^{(2)} x_g + \varepsilon_g^{(2)} \quad g = 1, 2, \dots, P$$

And

$$S_2 = \sum_{g=1}^p \lambda_g^{(2)} x_g$$

We ignore the intercepts and coefficients for S_1 in the model for each x . Therefore S_2 is the average of the regression coefficients for y on X controlling for S_1 .

We can now estimate a 2-component CCR model using S_1 and S_2 as predictors.

$$\hat{y} = \alpha^{(2)} + b_1^{(2)} S_1 + b_2^{(2)} S_2$$

In this new model S_1 and S_2 act as composite components of the vector of predictors X . Note that (a) we have re-estimated a new intercept and (b) the parameters $b_1^{(2)}$ and $b_2^{(2)}$ are the effects of respectively the first and second component S_1 and S_2 in the regression of y on S_1 and S_2 for the 2-component CCR model. We can decompose these into regularised effects for the regression of y on X using the fact that the components are linear combination of X :

$$\hat{y} = \alpha^{(2)} + \sum_{g=1}^P \lambda_g^{*(2)} x_g$$

Where $\lambda_g^{*(2)} = b_1^{(2)} \lambda_g^{(1)} + b_2^{(2)} \lambda_g^{(2)}$ is the regularised coefficient for x_g in the 2-component CCR model.

Step 3 – Estimation of remaining components 3..K

Continue this process for the remaining components. The K th component is defined as a weighted average of $\lambda_g^{(K)} x_g$ $g = 1, 2, \dots, P$

Where each $\lambda_g^{(K)}$ is estimated from the following 2-predictor model:

$$y = \alpha_g^{(K)} + \gamma_{1.g}^{(K)} S_1 + \gamma_{2.g}^{(K)} S_2 + \dots + \gamma_{(K-1).g}^{(K)} S_{K-1} + \lambda_g^{(K)} x_g + \varepsilon_g^{(K)} \quad g = 1, 2, \dots, P$$

And

$$S_K = \sum_{g=1}^p \lambda_g^{(K)} x_g$$

We ignore the intercepts and coefficients for S_1 through S_{K-1} in the model for each x . Therefore S_K is the average of the regression coefficients for y on X controlling for S_1 through to S_{K-1} .

We can now estimate a K -component CCR model using $S_1 \dots S_K$ as predictors.

$$\hat{y} = \alpha^{(K)} + b_1^{(K)} S_1 + b_2^{(K)} S_2 + \dots + b_K^{(K)} S_K$$

Where $S_1 \dots S_K$ act as composite components of the vector of predictors X . The parameters $b_j^{(K)}$ ($j = 1 \dots K$) are the effects of their respective components S_j in the regression of y on S_j for the k -component CCR model. We can decompose these into regularised effects for the regression of y on X using the fact that the components are linear combination of X :

$$\hat{y} = \alpha^{(K)} + \sum_{g=1}^p \lambda_g^{*(K)} x_g$$

Where $\lambda_g^{*(K)} = b_1^{(K)} \lambda_g^{(1)} + b_2^{(K)} \lambda_g^{(2)} + \dots + b_K^{(K)} \lambda_g^{(K)}$ is the regularised coefficient for x_g in the K -component CCR model.

Intuitively it can be seen that adding more components increases the variance of the resulting $\lambda_g^{*(j)}$ ($j = 1 \dots K$) whereas conversely fewer components decreases this variance and therefore impose greater regularisation on the coefficients.

Cross-Validation Procedure

Why Cross-validation?

The purpose of cross-validation (CV) is to investigate the extent to which a model estimated on a sample of cases predicts well for new cases. The sample on which the model is estimated is known as the “training” sample and the sample used to validate the model is known as the “validation” sample. Assessing the models performance on the validation sample is particularly useful for situations where the in-sample (training sample) model may over-fit the data, due to small sample size, high predictor correlations or where the number of potential predictors approaches the sample size. In such cases the usual model selection procedure of hypothesis testing in the training sample (F-test and t-tests) may give an inaccurate or misleading picture or may not provide useful information (due to lack of statistical significance).

In CCR a cross-validation procedure is required in order for us to select the optimal number of correlated components (K) and predictors (P) in the model. It can also be used to give an indication

of the importance the available predictors under the model. Given that most real data sets only have a finite (and often small) number of cases an efficient method of CV is required to ensure that we maximise both our training and validation samples.

M-fold validation

The method of cross-validation used for CCR and implemented in CORExpress is M-fold validation. Typically more than one run of M-fold validation is required. Each round provides one set of Cross-validated (CV) performance statistics. We can therefore think of each round as one observation for a sample (over all rounds of CV). This enables us to calculate a mean and standard error for each CV statistic. The usual rules of thumb for sampling apply, with a minimum of 50 rounds recommended where processing speed allows.

One round of M-fold validation randomly divides a sample of n cases into M mutually exclusive subgroups, known as folds. The M-fold validation process works as follows:

1. Create a mutually exclusive random partition of M continuous folds/subgroups on our sample of size n (we will assume in this example that the whole sample is to be used to develop the model which will be the case most of the time). This partition is selected such that we obtain roughly equal numbers of cases in each fold. M is typically selected to be a number between 5 and 10 (though any number between 2 and $n/2$ can be used) and it is a good practice to ensure that n is exactly divisible by M . We will assume for illustrative purposes that $n=100$ and $M=10$, so that each fold contained exactly 10 cases.
2. The first fold (in this case the first random selection of 10 cases) is held back as part of the (rolling) validation sample and the remaining 9 folds (90 cases) are used as training sample to estimate the model. The model's predictive performance is then used to score cases in the validation fold. For linear modelling we typically use CV R-squared to assess the performance of the model in the validation fold.
3. The result from (2) on fold 1 (CV R-squared for the model developed in absence of the first fold) is then stored and the process repeated for the second fold. So this second fold of 10 cases which formed part of the training sample in the previous step is now held back as an additional part of the validation sample and the model re-estimated on the remaining folds, which now act as the training sample. The model's predictive performance is then tested on this 2nd validation fold and the results aggregated with the previous CV performance (from fold 1).
4. The process then repeats for the remaining 8 folds.

Note that this process ensures that across all M -folds (in this case where $M=10$) :

- All folds are used as training sample exactly $9/10$ ($(M-1)/M$) times; All folds are used as validation sample EXACTLY ONCE.
- The implication of this is that the WHOLE SAMPLE IS RE-USED AS VALIDATION SAMPLE making this a very efficient method of cross-validation (All cases are the validation sample).
- We obtain M models with M sets of model parameters (this becomes more relevant when considering the CCR stepping down procedure in the next section)

For our linear modelling example we obtain a CV r-squared estimate which is an average of the CV r-squared estimates at each step (with each fold 1..M) excluded. This can be used to assess the models performance out-of-sample performance. Note that for Logistic Regression and Linear Discriminant Analysis (see last section) the algorithm uses CV Accuracy (% of cases correctly allocated in the CV sample) rather than R-squared to assess model performance.

Stepping Down Procedure

Variable selection is achieved using a simple stepping down procedure which works in conjunction with M-fold cross-validation. For a particular value of K (number of correlated components) we initially estimate a model with all possible predictors. This model is then evaluated using M-fold validation as described in the previous section to obtain a CV (Cross-validated) R-squared value (or Accuracy value in the case of Logistic Regression and Linear Discriminant Analysis). The model is then estimated on the entire (training) sample. The model coefficients are standardised. In the case of linear regression this involves multiplication of the unstandardized coefficients by the standard deviation of the predictor and division by the standard deviation of y.

The predictor with the lowest standardised effect in the entire training sample is then DROPPED and the process repeated again. This yields a CV R-squared for a new model minus the “least predictive” predictor. The model is estimated on the entire training sample and then we drop the predictor which has the lowest standardised coefficient. This process then repeats until we have CV R-squared values for every possible model with between 1 and the maximum number of possible predictors, unless a different range is specified by the user.

The CV performance for each possible model ranging from that with the maximum number of possible predictors down to a one predictor model is outputted by the program. The program automatically selects the optimal number of predictors on the basis of the model with the lowest CV R-squared for linear regression or Accuracy for Logistic/Discriminant Analysis. Having decided on the optimal number of predictors (via crossvalidation) the stepping-down procedure is then run again on the entire training sample to select the appropriate predictors.

Alternatively users can compare the CV performance for models with different number of predictors and use various rules-of-thumb to determine the number of predictors which represents the best trade-off between out-of-sample performance (CV R-squared/Accuracy) and simplicity (few predictors). The stepping-down procedure can then be re-run specifying the number of predictor required.

This procedure can be thought of as analogous to backward variable selection (as implemented in other regression programs) although the variable selection process is driven entirely by Cross-validation (CV) sample performance, rather than by hypothesis testing.

Final CCR Model

Note that the final regularised CCR model with optimal K and P is estimated on the entire training sample utilising all available information. The CV and Stepping-down procedure are means-to-an-end for selecting the optimal values of P (number of predictors) and K (number of components) and to provide additional guidance on predictor importance.

Extensions to Linear Discriminant Analysis and Logistic Regression

Linear Discriminant Analysis (LDA) and Logistic Regression are used when the dependent (y) variable is a dichotomous (two-category) group. These procedures are typically used when trying to predict whether a case is in or out of a particular group of interest. A classic example would be to predict who would vote for Party X in a general election.

A similar procedure is followed for these additional classes of models with appropriate model estimation procedure used at the core. For both LDA and Logistic Regression the primary CV performance measure used is Accuracy, which is defined as the proportion of cases correctly classified (given that probability of membership is greater than or less than or equal to 50%).

END

References

1. 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.
2. Magidson, J. (2010). Correlated Component Regression: A Prediction/Classification Methodology for Possibly Many Features. Proceedings of the American Statistical Association.
3. Magidson, J (2011). CORExpress® 1.0 Users Guide