A neural network based on canonical correlation for multicollinearity diagnosis

Jifu Nong
College of Science
Guangxi University for Nationalities
Nanning, China
njf93471@163.com

***Abstract*—We review a recent neural implementation of Canonical Correlation Analysis and show, using ideas suggested by Ridge Regression, how to make the algorithm robust. The network is shown to operate on data sets which exhibit multicollinearity. We develop a second model which not only performs as well on multicollinear data but also on general data sets. This model allows us to vary a single parameter so that the network is capable of performing Partial Least Squares regression to Canonical Correlation Analysis and every intermediate operation between the two. On multicollinear data, the parameter setting is shown to be important but on more general data no particular parameter setting is required. Finally, we develop a second penalty term which acts on such data as a smoother in that the resulting weight vectors are much smoother and more interpretable than the weights without the robustification term. We illustrate our algorithms on both artificial and real data.**

Keywords- Canonical correlation analysis; Roughness penalty; Multicollinearity; Partial least squares regression

I. INTRODUCTION

Canonical Correlation Analysis (CCA) is a statistical technique used when we have two data sets which we believe have some underlying correlation. Let us have sample vectors x_1 and x_2 drawn from two related data sets. Then in classical CCA, we attempt to find that linear combination of the variables which give us maximum correlation between the combinations. Let

$$y_1 = w_1^T x_1 = \sum_j w_{1j} x_{1j}$$

$$y_2 = w_2^T x_2 = \sum_j w_{2j} x_{2j}$$

Then we wish to find those values of w_1 and w_2 which maximize the correlation between y_1 and y_2. If the relation between y_1 and y_2 is believed to be causal, we may view the process as one of finding the best prediction of x_2 from the second data set by the sample, x_1 from the first data set and similarly of finding the best predictable criterion in the sample x_1 in order to predict the sample x_2.

Now it may be shown that a method of finding the canonical correlation directions is to solve the generalized eigenvalue problem

$$\begin{bmatrix} 0 & \sum_{12} \\ \sum_{21} & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \rho \begin{bmatrix} \sum_{11} & 0 \\ 0 & \sum_{22} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \quad (1)$$

where $\sum_{ij}$ is the covariance matrix of x_i and x_j.

It has recently been shown that solutions of the generalized eigenvalue problem

$$Aw = \lambda Bw \quad (2)$$

can be found using gradient ascent of the form

$$\frac{dw}{dt} = Aw - f(w)Bw \quad (3)$$

where the function $f(w): R^n \to R$.

Intuitively, what these criteria mean are that

1. The function is rather smooth.

2. It is always possible to find values of $w_i, i = 1,2$, large enough so that the functions of the weights exceed the greatest eigenvalue.

3. It is always possible to find values of $w_i, i = 1,2$ small enough so that the functions of the weights are smaller than the least eigenvalue.

4. For any particular value of $w_i, i = 1,2$, it is possible to multiply $w_i, i = 1,2$, by a scalar and apply the function to the result to get a value greater than the greatest eigenvalue.

5. Similarly, we can find another scalar so that, multiplying the w_i, by this scalar and taking the function of the result gives us a value less than the smallest eigenvalue.

6. The function of this product is monotonically increasing between the scalars defined in Eqs. (4) and (5).

A typical example of $f(\cdot)$ taken from Zhang and Leung (2000) would be $f(w) = \ln(w^T(t)w(t))$, which we use for all the experiments in this paper.

Using formulation (3) we have shown that the canonical correlation directions w_1 and w_2 may be found using

$$\frac{dw_1}{dt} = \sum_{12} w_2 - f(w) \sum_{11} w_1$$

$$\frac{dw_2}{dt} = \sum\nolimits_{21} w1 - f(w)\sum\nolimits_{22} w_2$$

Using the fact that, for zero mean data, $\sum_{ij} = E(x_i x_j^T)$, we derive the instantaneous versions

$$\Delta w_1 = \eta(x_1 y_2 - f(w) x_1 y_1)$$
$$\Delta w_2 = \eta(x_2 y_1 - f(w) x_2 y_2)$$

which was shown to provide a family of networks capable of performing CCA. If we use a single pair of outputs, y_1 and y_2 with corresponding weights, w_1 and w_2 the system of equations converges to the first generalized eigenvectors; if we wish subsequent correlation filters, we can use deflationary methods or other means of introducing some asymmetry to the learning rules. The theoretical analysis in Zhang and Leung (2000) shows that any function $f(\cdot)$ satisfying the above three criteria will cause convergence to the eigenvector with the greatest eigenvalue; we can confirm empirically that all functions we have used in the above system of equations have been successful in causing convergence to the canonical correlation directions.

II. The ridge model

The problem of multicollinearity arises in a regression problem whenever there is a linear dependency among the independent variables. That is, let $X = (x_0, x_1, \cdots, x_{p-1})$ where x_i is the $n \times 1$ vector of responses for the ith variable. The independent variables are said to have linear dependence whenever

$$\sum_{j=0}^{p-1} t_j x_j = 0 \tag{4}$$

for $t_j \neq 0$. To solve $X\beta = y$, the standard method is to multiply both sides of this equation by X^T and then solve for β to give $\beta = (X^T X)^{-1} X^T y$. If condition (4) holds then $(X^T X)^{-1}$ does not exist. Seldom does the above linear dependency actually hold, rather one nearly has linear dependency which implies that $(X^T X)^{-1}$ is ill-conditioned, hence any estimates using $(X^T X)^{-1}$ are poor.

Ridge regression is a popular method for dealing with multicollinearity within a regression model. The idea is fairly simple. Since the matrix $X^T X$ is ill-conditioned or nearly singular one can add positive constants to the diagonal of the matrix and ensure that the resulting matrix is not ill-conditioned. That is, consider the biased normal equations given by

$$(X^T X + kI)\beta = X^T y \tag{5}$$

where I is the identity matrix. This results in a biased estimate for b given by

$$\tilde{\beta} = (X^T X + kI)^{-1} X^T y \tag{6}$$

where k is called the shrinkage parameter. This has been shown to make the regression robust.

III. Application to CCA

The canonical correlation coefficient is given by

$$\rho = \frac{w_1^T \sum_{12} w_2}{(w_1^T \sum_{11} w_1)^{1/2} (w_2^T \sum_{22} w_2)^{1/2}} \tag{7}$$

which will clearly be difficult to calculate if the within class covariance matrices are singular or nearly so. Similarly, since the generalized eigenvectors found by solving

$$\begin{bmatrix} 0 & \sum_{12} \\ \sum_{21} & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \rho \begin{bmatrix} \sum_{11} & 0 \\ 0 & \sum_{22} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \tag{8}$$

may be equally well-defined as the eigenvectors found by solving

$$\begin{bmatrix} \sum_{11} & 0 \\ 0 & \sum_{22} \end{bmatrix}^{-1} \begin{bmatrix} 0 & \sum_{12} \\ \sum_{21} & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \rho \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \tag{9}$$

Vinod (1976) shows that the coefficients estimated from the usual CCA can be very unstable when the data are non-orthogonal, but after adding small constants to the diagonal of the correlation matrix of all variables before the usual CCA, a considerable improvement in the stability and reliability of regression coefficients is achieved. We also can consider this as kind of smoothing for the data; in Leurgans, Moyeed, and Silverman, a similar approach has been used to deal with functional data. Thus, from the above, we have good reasons to believe that the penalty term kI can make CCA more robust.

Now if we use $\sum_{11} + k_1 I$ and $\sum_{22} + k_2 I$ instead of $\sum_{11}$ and $\sum_{22}$ in our neural implementation, we get

$$\begin{bmatrix} 0 & \sum_{12} \\ \sum_{21} & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \rho \begin{bmatrix} \sum_{11} + k_1 I & 0 \\ 0 & \sum_{22} + k_2 I \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \tag{10}$$

Taking $w = (w_1^T \ w_2^T)^T$, we find the canonical correlation directions w_1 and w_2 using

$$\frac{dw_1}{dt} = \sum\nolimits_{12} w_2 - f(w_1)(\sum\nolimits_{11} + k_1 I) w_1$$
$$\frac{dw_2}{dt} = \sum\nolimits_{21} w_1 - f(w_2)(\sum\nolimits_{22} + k_2 I) w_2$$

We may propose the instantaneous rules

$$\Delta w_1 = \eta(x_1 y_2 - f(w_1) x_1 y_1 - f(w_1) k_1 w_1)$$
$$\Delta w_2 = \eta(x_2 y_1 - f(w_2) x_2 y_2 - f(w_2) k_2 w_2)$$

This algorithm, in fact, does perform an approximation to CCA but we have found experimentally that it does slow learning. The k_i parameters are optimal when rather large but note that this has a detrimental effect in that it affects both the second term and also the first term which causes growth towards the canonical correlation vectors. To restrict this effect, we restructure Eq. (10) to get

$$\begin{bmatrix} 0 & \sum_{12} \\ \sum_{21} & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \rho \begin{bmatrix} (1-k_1)\sum_{11}+k_1 I & 0 \\ 0 & (1-k_2)\sum_{22}+k_2 I \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \qquad (11)$$

Thus the update rule for the weights is

$$\begin{aligned} \Delta w_1 &= \eta(x_1 y_2 - f(w_1)(1-k_1)x_1 y_1 - f(w_1)k_1 w_1) \\ \Delta w_2 &= \eta(x_2 y_1 - f(w_2)(1-k_2)x_2 y_2 - f(w_2)k_2 w_2) \end{aligned} \qquad (12)$$

This method of weight change is the first innovation in this paper. We may however consider generalizing this method by using different bias-inducing terms. For example, we may wish to produce smoothly changing CCA parameters and so we may wish to introduce a term which penalizes roughness in the CCA weights.

Functional data analysis (FDA) has been developed for analyzing functional data. In FDA, we treat the data as consisting of functions not of vectors. We take samples at time points $t_1, t_2, \cdots$ and regard $\{x(t_j), j=1,2,\cdots\}$ as multivariate observations. In this sense the original functional $x(t)$ can be regarded as the limit of $\{x(t_j)\}$ as the sampling interval tends to zero and the dimension of multivariate observations tends to infinity. The central idea of doing FDA is using a roughness penalty to incorporate smoothing. The most popular measure of roughness is the second derivative of the function form, i.e. a measure of the rapidity of the variability of the function f is given by

$$R(f) = \int (f\,''(x))^2\, dx \qquad (13)$$

Since we do not care about the sign of the roughness, only its magnitude, we define a penalty for roughness by

$$PEN_2(f) = \| D^2 f \|^2 = \int f(t) D^4 f(t) dt \qquad (14)$$

where D^2 is the second derivative operator and D^4 is the fourth derivative operator.

The phrasing of the parameters in this way allows us to consider a family of solutions found by varying the magnitude of the k_i parameters. For example, if $k_1 = k_2 = 1$ in Eq. (12), we revert to $Aw = \lambda Bw$, gives us the solution for Partial Least Square (PLS):

$$\begin{bmatrix} 0 & \sum_{12} \\ \sum_{21} & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \rho \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \qquad (15)$$

The PLS regression method has been used extensively, specially for calibration tasks in chemometrics.

It may be helpful to compare PLS with the CCA, which is an maximization of the correlation between a similar pair of scores. Recall that PLS maximizes covariance, not correlation. Interpreting the coefficients of canonical variates requires the usual stringent assumptions underlying multiple regression of either canonical variate upon the variables of the other block. Such assumptions are unlikely to obtain when predictors or outcomes are intentionally redundant. In contrast, by maximizing covariance between the latent variable scores, PLS optimizes the usefulness of the analysis for subsequent studies of intervention.

IV. EXPERIMENTS

A. *Multicollinear data*

We generate two artificial datasets x_1 and x_2

$$x_1 = (x_{11}, x_{12}, \cdots, x_{1n}), x_2 = (x_{21}, x_{22}, \cdots, x_{2n})$$

in which we take $n = 20$; each x_{1n} is linear combination of b_1 and each x_{2n} is linear combination of b_2 where

$$b_1 = (b_{11}, b_{12}, \cdots, b_{1p}), b_2 = (b_{21}, b_{22}, \cdots, b_{2p})$$

with *p*=4. Now, we have two data sets, each of which has very high internal correlations. Now we create a strong correlation between the two data sets by defining

$$c = (x_{11} + x_{21}) / 2 \qquad (16)$$

and then set $x_{11} = x_{21} = c$. Now, the first elements in both data sets are exactly same, and each of these new first elements has a high correlation with other internal elements and also has a correlation with elements of the other data set. These provide the only correlations between the two data sets. We use two algorithms on this data set: one is our new algorithm (13) with the smoothing parameter, we see that the existing neural algorithm has had a great deal of difficulty with this data set while the new algorithm (13) has identified the major correlations very effectively.

TABLE I. WEIGHTS VALUE OF THE MULTICOLLINEARITY DATA

	Existing algorithm		**New algorithm**	
	1	***2***	***1***	***2***
1	20.6918	20.5765	21.6644	21.6222
2	20.0453	0.5527	0.0022	20.0235
3	20.8561	0.946	20.0045	0.0048
4	0.4852	20.2	0.0031	20.029
5	20.5817	20.1928	0.0016	20.0112
6	0.5947	20.0139	0.0001	0.0239
7	20.0364	0.6404	20.0039	0.0117
8	20.0373	20.1642	20.0006	20.0059
9	0.1009	20.7506	20.0022	0.0037
10	0.4595	20.2232	20.0037	0.0057
11	20.0454	0.0366	0.0042	0.0169
12	20.2113	20.1909	0.0001	20.0071
13	0.0907	0.1842	0.0022	0.0051
14	20.0137	0.3383	20.0015	0.0007
15	20.299	0.4471	0.0038	20.0059

B. *Children's gait data*

The Children's Gait Data has been used in Leurgans et al. collected by the Motion Analysis Laboratory at the Children's Hospital, San Diego, CA,. The data set consist of the angular rotations in the sagittal plane of the hip and knee of 39 normal 5-year-old children. The observations are taken over a gait cycle consisting of one double step taken by each child, and time is measured in terms of the cycle which has been discretized to a regular grid of 20 points.

Fig. 1 shows the results of the simulation in terms of the weight parameters with *k*=0.9 while Fig. 1 shows the results using a previous neural algorithm. It is clear that Fig. 1 is rather difficult to interpret while Fig.1 is much more interpretable. Also the smoothness of Fig.1 gives us somewhat greater confidence in the predictive power of this result since Fig. 1 appears to be a noisy solution. Because we are not interested in this specific data, we do not analyse the experiment's results, but from Fig.1, we can see the hip curve in the middle of the cycle occurs a little later than that in the knee curve, which concurs with the interpretation in Leurgans et al. Figs.1 shows the first and second weights value from the algorithm with the roughness penalty smoothing term (11) and (14). Both the first and second weights' values could be transformed roughly to being identical for the hip and the knee by speeding up the hip cycle relative to the knee cycle in the first half of the cycle and slowing it down in the second. Since the main interest is in comparing the curves, all of the weights value shown in Figs.s has been normalized so that the integral of their squares is equal to 1.

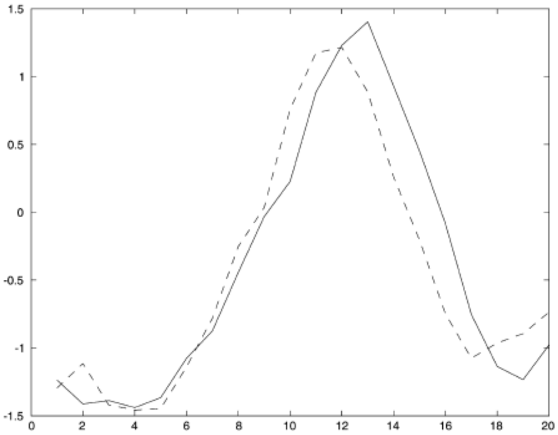

Figure 1. Canonical variate weights using New learning rule, the solid line for hip, dash line for knee.

We also found experimentally that, if we use the simple ridge solution (11), we need to use a large *k*, (20–50), which has an adverse effect on the growth part of the algorithm and causes decay away from the optimal directions. The final result is that the estimate of canonical correlation is very poor. If we use the hybrid solution (13), we just need a number between 0 and 1.

V. Conclusion

We have used the basic idea from ridge regression to create an algorithm which is robust with multicollinear data. We have shown the effectiveness of the algorithm on artificial data which was designed to be multicollinear and on a real data set which has a limited number of samples in relation to its dimensionality. With a second real data set-one which does not exhibit multicollinearity--we have shown that the addition of the robustification parameter does not materially affect the results for a wide range of parameter values.

We have introduced a penalty term which penalizes roughness in the canonical correlation directions and shown that the resulting vectors are much more interpretable than the original. The resulting canonical correlation vectors are more suited to prediction than those achieved without the penalty term.

Finally, we have used lateral connections to find more than one canonical correlation vector.

Acknowledgment

The work was supported by Guangxi department of education scientific research project under Grand No. 201204LX083.

References

[1] Diamantaras, K. I., & Kung, S. Y.. Multilayer neural networks for reduced-rank approximation. IEEE Transactions on Neural Networks, 1994, 5(5), 684–697.

[2] Fyfe, C.. Introducing asymmetry into interneuron learning. Neural Computation, 1995, 7(6), 1167–1181.

[3] Geladi, P., & Kowalski, B.. Partial least squares regression: a tutorial. Analytica Chimica Acta, 1986, 185, 1–17.

[4] Gou, Z., & Fyfe, C.. A family of networks which perform canonical correlation analysis. International Journal of Knowledge-based Intelligent Engineering, 2001, 5(2), 76–82.

[5] Horel, A. E., & Kennard, R. W.. Ridge regression: biased estimation for nonorthogonal problems. Technometrics, 1970, 12, 56–67.

[6] Hoskuldsson, A.. PLS regression methods. Journal of Chemometrics, 1988, 2, 211–228.

[7] Lai, P. L., & Fyfe, C.. A neural network implementation of canonical correlation analysis. Neural Networks, 1999, 12(10), 1391–1397.

[8] Leurgans, S. E., Moyeed, R. A., & Silverman, B. W.. Canonical correlation analysis when the data are curves. Journal of the Royal Statistical Society, Series B, 1993, 55(2), 725–740.

[9] Olshen, R. A., Biden, E. N., Wyatt, M. P., & Sutherland, D. H.. Gait analysis and the bootstrap. The Annals of Statistics, 1989, 17, 1419–1440.

[10] Ramsay, J. O., & Silverman, B. W.. Functional data analysis. Berlin: Springer, 1997.

[11] Zhang, Q., & Leung, Y. W.. A class of learning algorithms for principal component analysis and minor component analysis. IEEE Transactions on Neural Networks, 2000, 11(1), 200–204.