Online Journal of Applied Knowledge Management
A Publication of the International Institute for Applied Knowledge Management

Volume 5, Issue 2, 2017

Multivariate text mining for process improvement using cross-canonical correlation analysis

Jose Luis Guerrero Cusumano, Georgetown University, McDonough School of Business, guerrerj@georgetown.edu

Abstract

Text analysis is a useful tool to determine what a company and its customers want in order to improve processes and methodologies of analysis. Searches in databases may have a time series component that determines the importance and sequences of multivariate searches and its structure. This paper presents a methodology to simplify and model multivariate searches in time using the Canonical Correlation approach. The techniques shown provide a robust methodology to simplify the analysis and create predictive models taking into account temporal dependencies.

Keywords: Text analysis, Google correlate, multivariate time series, cross correlation, canonical correlation, Radic matrices and determinant, predictive modelling.

Introduction

Big data and unstructured data can give a deeper and more accurate understanding into business. If 20% of the data available to enterprises is structured data, the other 80% is unstructured. Unstructured data — everything from social media posts and sensor data to email, images, and web logs — is growing at an unprecedented pace. Twitter sees over 175 million tweets each day and has more than 300 million active users; 571 new websites are created every minute of every day. Moreover, the world creates 2.5 quintillion bytes of data per day from unstructured data sources like sensors, which are fundamental for process improvement. We generate unstructured data incessantly from internal company texts, social media data, mobile data, and website content.

Text mining is essential and provides deep analysis of text-based documents. The problem is that the practitioner finds many analytics, statistics, and linguistics concepts very difficult to deal with, and it is compounded by additional difficulties inherent to the user’s native language and culture. The word frequency distribution in text in any language follows different statistical patterns and it is difficult to ascertain.

This paper looks at a different approach to text mining and concentrates on the temporal relationships of the ranking of frequency of words in order to create a predictive model. In practice, words and expressions found in web searches may be indicators of underlying processes to be studied and determined. A typical example is the following: let X and Y be vectors of words in web searchers where it is assumed that X precedes Y, where vector X can be [fever, cough, weakness], and vector Y can be [flu, pneumonia, cancer] (Ginsberg, 2008). In this previous example, the aim is to create a temporal predictive model for Y based on vector X. The model is temporal because it is suspected that vector X [fever, cough, weakness] may precede vector Y [flu, pneumonia, cancer].

The standard approach for multivariate time series with co-integration, the vector autoregressive moving-average model with exogenous variables, is called the VARMAX(r,l,s)(p,q) model. The form of the model can be written as

$$Y_t = \sum_{i=1}^{r} \Phi_i Y_{t-i} + \sum_{i=0}^{s} \Xi_i X_{t-i} + \varepsilon_i - \sum_{i=0}^{l} \Theta_i \varepsilon_{t-i}$$

Where the output variables of interest, vector Yt= (y1, y2,…, yp)t can be influenced by other input variables, Xt= (x1, x2,…, xq)t which are determined outside of the system of interest. The variables Yt are referred to as dependent, response, or endogenous variables, and the variables X_t are referred to as independent, input, predictor, regressor, or exogenous variables.

The unobserved noise variables, □$_t$, are a vector white noise process. The number of parameters of this complex model is at least the product of r•l•s•p•q. As an example, for a simple model of four dependent variables and three independent variables, which has a simple ARMA(2,2) will require 48 parameters to estimate and deploy the model. Tiago (1989) warned about the proliferation of hyper-complex models that were not able to take advantage of the high correlation among the variables of vector Yt= (y1, y2,…, yp)t in Vector ARMA (VARMA) models (Doornik, 1996).

In this paper, an alternative approach is proposed to minimize the complexity of the problem at hand. In other words, to create a more general, succinct model that takes into account the correlations between Yt= (y1, y2,…, yp)t and variables, Xt= (x1, x2,…, xq)t and their cross-correlations using a Canonical Correlation approach. Two main tools are combined: Multivariate analysis via Canonical Correlation Analysis and Time Series via Cross Correlation Analysis applied to text Analysis for process improvement and prediction. As an example, the multivariate data from Google correlate will be used.

Google Correlate and Spurious Correlations

Google Correlate (Mohebbi, 2011) finds queries that are correlated with other queries or with user-supplied data across time (Stephens-Davidowitz & Hal Varian, 2015). Google Correlate provides Pearson correlation coefficients among the word searches. Google Correlate will provide up to 100 correlates. Some of these correlation will be spurious and others significant for this research (Haig, 2006). Consider that the research is able to separate spurious and not spurious relationships (Ginsberg, 2008).

There are different ways of determining spurious correlation. The first one is the common sense and knowledge of the researcher. The second one is based on statistics, namely, first two time series variables are regressed on each other and get a large coefficient of determination (Chatfield, 2005). Granger and Newbold (1974) suggested that if the coefficient of determination is larger than the Durbin-Watson statistic, then the researcher may worry about spurious regression. Spurious correlation may occur when two random walks are observed with a positive drift. Co-integration is another approach that is used later in this paper. On the non-spurious correlation, the researcher should determine which queries could become dependent queries and independent ones to create a multivariate time series predictive model.

Methodology Using Canonical Correlation

Consider the dependent variable query vector Y= (y1, y2,…, yp) and X= (x1, x2,…, xq) as an independent variable query at one moment in time. Often in practice one vector of variables is a criterion set and the other vector of variables is a predictor set. Let us consider the covariance and correlation structure of the variables Y and X.

With covariance structure $\Sigma = \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix}$ and with correlation structure $R = \begin{bmatrix} R_{XX} & R_{XY} \\ R_{YX} & R_{YY} \end{bmatrix}$.

What is Canonical Correlation? Interrelationships between sets of multiple independent variables and multiple dependent measures quantify the strength of the relationship. Objectives of Canonical Correlation Analysis are to determine relationships among sets of variables, achieve maximal correlation, and explain the nature of relationships among sets of variables. Canonical correlation will be used to find linear/nonlinear combinations of both sets of variables Y and X that are maximally correlated. The objective in canonical correlation analysis is to determine simultaneous relationships between the two sets of variables. The canonical correlations between X and Y can be found by solving the eigenvalues equations (Konishi, 1979):

$$\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}\widehat{w}_X = \rho^2\widehat{w}_X$$

$$\Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}\widehat{w}_Y = \rho^2\widehat{w}_Y$$

Where the eigenvalues $\square^2$ are the squared canonical correlations and the eigenvectors W_x and W_y are the normalized canonical correlation basis vectors. The canonical eigenvectors summarize the correlation information between the X and Y variables.

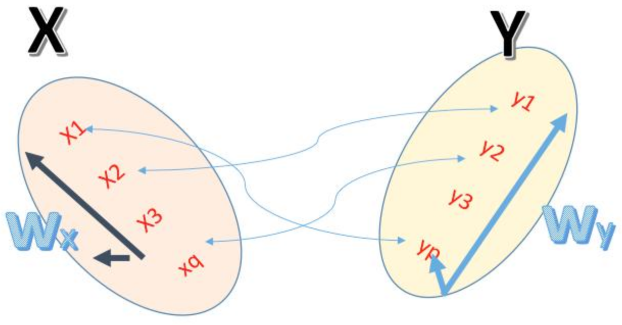


Figure 1. Canonical Correlation Analysis Approach

Vectors W_x and W_y are called the canonical correlation vector, which are the transformation of the original vector Y and X. The determinant of square matrix could be considered a total measure of correlation. In the case of two variables y and x, we have the following expression:

$\left|R_{yx}\right| = 1 - \rho_{yx}$, therefore, for two variables to be uncorrelated is equivalent that $\left|R_{yx}\right| = 1$

In general, the determinant of the correlation matrix will equal 1.0 only if all correlations equal 0, otherwise the determinant will be less than 1. The determinant is related to the volume of the space occupied by the swarm of data points represented by standard scores on the measures involved. When the measures are uncorrelated, this space is a sphere with a volume of 1. When the measures are correlated, the space occupied becomes an ellipsoid whose volume is less than 1 (Konishi, 1979).

The ideal situation for canonical correlation to be most effective, in the sense of having smaller number of canonical vectors in order to reduce the dimensionality of our problem (consisting of q-variate vector Y & p-variate vector X), would be that the determinants for the dependent and independent have the following properties (Konishi, 1979):

$$\left|R_{XX}\right| \approx 1, \left|R_{YY}\right| \approx 1$$

Namely, those components of vector Y (& vector X) are not highly correlated to each other. This assures the stability of the coefficients for vectors W_x and W_y. Specifically, for the reduction of dimensionality, each of the components of vector Y should be highly correlated to components of vector X.

Given that R_{YX} is not a square matrix, the usual definition of determinant of a matrix cannot be applied. However, we can use the definition of non-square matrix determinant given by Radic (1969), namely, let A= (a_{ij}) be an mXn matrix with m =< n. The determinant of A is defined as

$$|A| = \sum_{1 \le j_1 \le \ldots \le j_m \le n} (-1)^{r+s} \det \begin{vmatrix} a_{1j_1} & \ldots & a_{1j_m} \\ \ldots & \ldots & \ldots \\ a_{mj_1} & \ldots & a_{mj_m} \end{vmatrix}$$

where $j_1, \ldots j_m \in N; r = 1 + 2 + \ldots + m; s = j_1 + \ldots + j_m$,

As an example, for a rectangular matrix with 2 rows by 3 columns (2 by 3), namely A = [A_1, A_2, A_3] , its determinant is given by:

$$\det \begin{vmatrix} a_1\, a_2\, a_3 \\ b_1\, b_2\, b_3 \end{vmatrix} = \det \begin{vmatrix} a_1\, a_2 \\ b_1\, b_2 \end{vmatrix} - \det \begin{vmatrix} a_1\, a_3 \\ b_1\, b_3 \end{vmatrix} + \det \begin{vmatrix} a_2\, a_3 \\ b_2\, b_3 \end{vmatrix}$$

As in the case of square matrices, the determinant of rectangular matrices represents the volume generated by its column. Namely, every real m x n matrix A = [A_1, ..., A_n] determines a polygon in R^m (the columns of the matrix correspond to the vertices of the polygon) and vice versa. First of all, the determinant of the rectangular matrix [A_1, . . .,A_n] is related to a volume of the polyhedron (A_1,..A_n). In general, the determinant of a m x (m +1) matrix [A_1, . . . , A_{m+i}] is proportional with a volume of the orientated m-simplex (A_i . . . A_{m+1}), as well (Sušanj & Radic, 1994):

$$\left|A_1, A_2, \ldots, A_{m+1}\right| = m! Vol\left(A_1, A_2, \ldots, A_{m+1}\right)$$

Radic (1969) showed if a row of A is identical to some other row or is a linear combination of other rows then |A| =0 and therefore the volume generated will be equal to zero, which is an indication of high collinear variables. In our case, the ideal situation for reduction of dimensionality is that (Stanimirović, 1997)

$Radic\left|R_{YX}\right| \approx 0$, namely, that vectors Y and X are highly correlated.

In short, the conditions for stability of the canonical coefficient and the maximum reduction of dimensionality would be

$$\left|R_{XX}\right| \approx 1, \left|R_{YY}\right| \approx 1, Radic\left|R_{YX}\right| \approx 0$$

In the appendix, a test to determine whether $\left|R_{XX}\right| \approx 1, \left|R_{YY}\right| \approx 1$ is provided.

Time Series Approach with Canonical Correlation for Google Correlate

The concept of cross-correlation relates to the correlation of two time series or vectors Y and X at different points in time. As a simple example of cross-correlation, consider the problem of determining possible leading or lagging relations between two time series x_t and y_t.

If the model $y_t = Ax_{t-l} + w_t$ holds, the series x_t is said to lead y_t for $l > 0$ and is said to lag y_t for $l < 0$. This assumes that noise w_t is uncorrelated with the x_t series. Hence, the analysis of leading and lagging relations might be important in predicting the value of y_t from x_t (Shumway, 2011).

In general, the vectors X and Y (or alternatively the canonical correlation vectors W_x & W_y) are considered time dependent, namely

$$Y_t = \begin{bmatrix} y_{1t} \\ \ldots \\ y_{qt} \end{bmatrix}, X = \begin{bmatrix} x_{1t} \\ \ldots \\ x_{pt} \end{bmatrix}, or, W_{Y_t} = \begin{bmatrix} W_{y_{1t}} \\ \ldots \\ W_{y_{\min(q,p)t}} \end{bmatrix}, W_x = \begin{bmatrix} W_{x_{1_t}} \\ \ldots \\ W_{x_{\min(q,p)_t}} \end{bmatrix}$$

The Cross Correlation matrix can be defined as a function between vectors X(t) and Y(t) (or $W(t)_x$ & $W(t)_y$). The cross correlation function (CCF) is helpful for identifying lags of the x-variable that might be useful predictors of y_t· Namely, the values of the W_x-variable will be used to predict future values of W_y. In the relationship between two time series (W_{yt} and W_{xt}), the series W_{yt} may be related to past lags of the W_{xt}-series.

The cross correlation matrix ρ for this canonical correlation problem is defined as $\rho = \left[CCF_{W_i y_{(t+k)}, W_i xt} \right]$ where CCF is defined by:

$$CCF_{Wy_{(t+k)}, Wxt} = \frac{\sum_{t=1}^{n-k} \left(W_{xt} - \bar{W}_x \right) \left(W_{y,t+k} - \bar{W}_y \right)}{S_{Wyt} S_{Wxt}}$$

Observe that the matrix ρ takes into account only the cross correlation between the dependent and independent canonical correlation variables W_{yi} and W_{xi}. For testing (Shumway, 2011) the cross correlation function (CCF), have the following asymptotic approximation, namely, the large sample distribution of CCF(k), for k = 1, 2, . . . , K, where k is fixed but arbitrary, is normal with mean zero and standard deviation $n^{-\frac{1}{2}}$ if at least one of the processes is independent white noise. Therefore, a simple test for the hypothesis of no cross-correlation, $H_0)\,\rho_{y_t,x_{t-k}}=0, H_0)\,\rho_{y_t,x_{t-k}}\neq 0$ at α = 5% is to whether the sample cross correlation $r_{y_t,x_{t-k}}$ is included in confidence interval $-\frac{2}{\sqrt{n}}<r_{y_t,x_{t-k}}<\frac{2}{\sqrt{n}}$ to fail to reject $H_0)\,\rho_{y_t,x_{t-k}}=0$ or alternatively for the canonical correlation analysis is: $-\frac{2}{\sqrt{n}}<CCF_{Wy_{(t+k)},Wxt}<\frac{2}{\sqrt{n}}$

The CCF is the first step to create a predictive model taking into account the multiple correlations between vector X and Y. Combining the two approaches will obtain the following graphical summary:

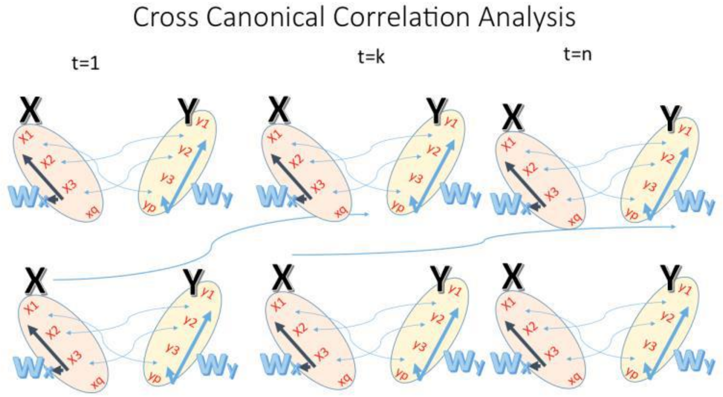


Figure 2. Canonical Analysis with a Time Series Approach

Without loss of generality, consider the case vectors X and Y are highly related to each other and within components of vectors X and Y have small correlations, namely $|R_{XX}|\approx 1, |R_{YY}|\approx 1, Radic|R_{YX}|\approx 0$; then a small number of canonical correlation variables (W_x , W_y) will explain the relationship between the X and Y vectors.

In the relationship between two time series (W_{yt} & W_{xt}), the series $W_{yt\text{ may}}$ be related to past lags of the W_{xt}-series. Consider the first canonical correlation variables $\left(W_{1,X_t}, W_{1,Y_t}\right)$ with cross correlation $CCF(k),\ for\ k\ =\ 0,\pm 1,\ \pm 2,\ ...,\ \pm K$

Let CCF_{max} be the largest of CCF(k) for a particular k^*, which is significant at an α level; then the suggested model for the lagged relationship would be $W_{1,Y_t}=AW_{1,X_{t-k^*}}+\varepsilon_t$. For example, if CCF_{max} for k=1, then the correlation between W_{1,Y_t} is measured in time period t and $W_{1,X_{t-1}}$ in time period t-

1, i.e. the correlation between W_{1,Y_t} is looked at in a time period and $W_{1,X_{t-1}}$ in the previous time period.

In the case that $|R_{XX}| \approx 0, |R_{YY}| \approx 0, Radic|R_{YX}| \approx 0$, a similar approach can be used considering Principal Components Analysis (PCA) (Bilodeau & Brenner, 1999). PCA will be explained in a following section below.

Collinearity, Canonical Correlation and Google Correlate

The case for $|R_{XX}| \approx 0$ or $|R_{YY}| \approx 0$

Given the nature of Google Correlate, which finds queries highly related to each other, the hypothesis to test $|R_{XX}| = 1$ or $|R_{YY}| = 1$ may be rejected. In practice, Google Correlate may have values supporting $|R_{XX}| \approx 0$ or $|R_{YY}| \approx 0$. These correlation matrices for vectors Y and X may be singular, i.e., the variables of vectors Y and X may be collinear.

A square matrix is singular if and only if its determinant is 0. This near singularity of the covariance matrix Σ_{XX} and Σ_{YY} will create instability in the canonical vector coefficient,

$$\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YY}^{-1}\boldsymbol{\Sigma}_{YX}\widehat{w}_X = \rho^2\widehat{w}_X$$
$$\boldsymbol{\Sigma}_{YY}^{-1}\boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}\widehat{w}_Y = \rho^2\widehat{w}_Y$$

In practice, Σ_{XX} and Σ_{YY} will always have some degree of singularity. The researcher needs to determine his/her tolerance to correlation among the variables of vectors Y and X.

If $|R_{XX}|$ and $|R_{YY}|$ are close to 0, we can take advantage of this situation using firstly Principal component analysis (PCA) on the vectors Y and X, namely Z_Y= (Zy_1, Zy_2,…, Zy_q) and Z_X= (Zx_1, Zx_2,…, Zx_q) and then apply canonical correlation analysis, (Jolliffe, 2002). Principal component analysis is simply an orthogonal transformation of the original Y into Z_Y and X into Z_X.

Canonical correlations are invariant with respect to affine transformations of the variables. An affine transformation is simply a translation of the origin followed by a linear transformation. In mathematical terms an affine transformation of R_n is a map from R_n to R_n of the form (Borga, 2001):

$F(p) = Ap + q$ for every vector p belonging to R^n

where A is a linear transformation of R^n and q is a translation vector in R^n, therefore the application of principal components to the vectors Y and X will not alter the results of canonical correlation analysis.

Without loss of generalization, consider a simple bivariate problem with x_1, x_2 and y_1, y_2. A number of authors (Preisendorfer & Mobley, 1988) have suggested that there are advantages in calculating principal components z_{x1}, z_{x2} separately for x_1, x_2 and z_{y1}, z_{y2} separately for y_1, y_2, then performing the canonical correlation analysis on z_{x1}, z_{x2} and z_{y1}, z_{y2} separately rather than x_1, x_2 and y_1, $y_{2.}$

Also, z_{x1}, z_{x2} and z_{y1}, z_{y2} are exact linear functions of x_1, x_2 and y_1, $y_{2.}$, respectively, and, conversely, x_1, x_2 and y_1, $y_{2.}$ are exact linear functions of , z_{x1}, z_{x2} and z_{y1}, z_{y2}, respectively.

Consider the new equation for the q-variate Y and p-variate X whose values were transformed by Principal components Z_Y= (Zy_1, Zy_2,…, Zy_q) and Z_X= (Zx_1, Zx_2,…, Zx_q), namely

$$\Sigma_{Z_{XX}}^{-1}\Sigma_{Z_{XY}}\Sigma_{Z_{YY}}^{-1}\Sigma_{Z_{YX}}\widehat{w}_{Z_X} = \rho^2\widehat{w}_{Z_X}$$

$$\Sigma_{Z_{YY}}^{-1}\Sigma_{Z_{YX}}\Sigma_{Z_{XX}}^{-1}\Sigma_{Z_{XY}}\widehat{w}_{Z_Y} = \rho^2\widehat{w}_{Z_Y}$$

By construction of the principal components, we have that $\Sigma_{Z_{YY}}$ and $\Sigma_{Z_{XX}}$ are diagonal matrices and $\left|R_{Z_{XX}}\right| = \left|R_{Z_{YY}}\right| = 1$

The number of principal components will be exactly q and p respectively, namely Z_Y= (Zy_1, Zy_2,…, Zy_q) and Z_X= (Zx_1, Zx_2,…, Zx_q). Some researchers may be tempted to drop some of the principal components (the least informative ones) at this stage before performing canonical correlation analysis. In general, this is not advisable, as some Zx_j may be dropped which are highly related to Zy_j. The researcher should use all principal components to perform the canonical correlation analysis except when the number p and q are extremely large to reduce the dimensionality of the problem.

Similarly as in the previous section, the series $W_{zyt\ may}$ be related to past lags of the W_{zxt}-series. Consider the first canonical correlation variables $\left(W_{1,ZX_t}, W_{1,ZY_t}\right)$ with cross correlation $CCF\left(k\right)$, *for* $k = 0, \pm 1, \pm 2, \ldots, \pm K$

Let CCF_{max} be the largest of CCF(k) for a particular k^* which is significant at an α level, then the suggested model for the lagged relationship would be $W_{1,ZY_t} = AW_{1,ZX_{t-k^*}} + \varepsilon_t$.

In some situations like image processing and analysis, we may have p and q >> n, then a reduction technique such as PCA should be use to transform p and q to smaller set of variables with dimension p' and q' in such a way that p' and q' << n which could be attained. When p and q >> n and hypotheses $\left|R_{XX}\right|=1, \left|R_{YY}\right|=1$ is rejected, it is then advisable to perform PCA on p and q variables separately reducing number of variables to smaller set p' and q', then apply canonical correlation analysis to these reduced number of principal components.

Application: Process Improvement with Google Correlate

The text mining application will be based on the information provided by Google Correlate. Google Correlate is a tool on Google Trends which enables finding queries with a similar pattern to a target data series. The target can either be a provided real-world trend (e.g., a data set of event counts over time) or an entered query. Google Correlate uses web search activity data to find queries with a similar pattern to a target data series. Let X and Y be vectors of words where it is assumed that X precedes Y, namely in the classical Google Correlate example of the flu, vector X can be fever, cough, weakness and vector Y can be flu, pneumonia, etc (Ginsberg, 2008).

Consider X(t) and Y(t) Google Correlate time dependent searches and W_{yt} and W_{xt} their canonical correlation counterparts. The cross correlation function will be used to determine the level of lag dependence between W_{yt} and W_{xt} to create a predictive model.

This methodology will be applied to the Toyota sudden acceleration problem. Since 1999, at least 2,262 Toyota and Lexus owners have reported to the National Highway Traffic Safety Administration, the media, the courts and to Safety Research & Strategies that their vehicles have accelerated suddenly and unexpectedly in a variety of scenarios. These incidents have resulted in 815 crashes, 341 injuries and, 19 deaths potentially related to sudden unintended acceleration. Figure 3 explains the order of Toyota acceleration and accident problems.

2006	**Sept.** National Highway Traffic Safety Administration (NHTSA) opens investigation on reports of "surging" in Camrys: closes investigation a year later	**2010**	**Jan. 21** Toyota recalls 2.3 million vehicles to correct separate problems that could cause gas pedal to stick Jan 26 Toyota suspends sales: halts production of eight models due to gas pedal recall; the next day. adds 1.1 million to floor mat recall
2007	**March** Toyota receives reports about accelerator pedal glitch in Tundra truck Sept. Toyota recalls some Lexus and Camry models to secure floor mat that could trap gas pedal, cause acceleration		**Feb. 1** Toyota says it had developed fix for sticking gas pedal; begins shipping part to dealers **Feb. 3** NHTSA says it has received more than 100 complaints about brake problems in Prius hybrids
2008	**Jan.** NHTSA investigates unintended acceleration in Toyota Tacoma pickups: probe closed in Aug. after no detect found		**Feb. 4** Toyota says gas pedal recall could cost $2 billion; total recall is 8.1 million; says Prius problem is software glitch; NHTSA opens Prius probe
2009	**Aug. 28** Off-duty Calif. Highway Patrol officer, family killed after gas pedal in Lexus is caught under floor mat **Sep. 29** Toyota issues recall for 3.8 million vehicles due to risk of gas pedal becoming caught under floor mat Nov. 4 NHTSA accuses Toyota of providing owners with "inaccurate and misleading information" about floor mat recall **Nov. 25** Toyota recalls at least 4 million vehicles to reconfigure gas pedals		**Feb. 5** Toyota CEO makes first public appearance to apologize for recall problems **Feb. 9** 437.000 Priuses recalled for brake problems: NHTSA says it has received complaints about steering problems in Corollas **Feb. 22** Federal prosecutors open criminal investigation into Toyota's safety problems **Feb. 23** Congress begins hearings on Toyota recalls

Figure 3. Events related to Toyota accidents and Recall

The predictive model created helps to determine the number of lags (weeks) which could have been used for the detection of Toyota Accidents.

Multivariate Approach using Canonical Correlations

Let us consider a vector approach with canonical correlation analysis. Using Google Correlate two set of variables were chosen from January 4th 2004 to January 17th 2010, namely vector Y =

(Toyota accident, Toyota crash, Toyota problem) and vector X = (Sudden acceleration, Acceleration problem). The following Table 1 shows the correlation among the Y and X vectors which are significant at $\alpha = 1\%$.

Table 1. Correlation among the Vector Y variables and Vector X variables

Row	**Toyota accident**	**Toyota Crash**	**Toyota problem**	**Sudden acceleration**	**Acceleration problem**
Toyota Accident	1.000	0.691	0.360	0.455	0.498
Toyota Crash	0.691	1.000	0.218	0.314	0.316
Toyota Problem	0.360	0.218	1.000	0.243	0.325
Sudden Acceleration	0.455	0.314	0.243	1.000	0.565
Acceleration Problem	0.498	0.316	0.325	0.565	1.000

Let's consider first the correlation between vector X and Y. The following table shows the correlation among the Y and X vectors which are significant at family $\alpha = 1\%$, (using a Bonferroni adjustment).

Table 2. Correlation between dependent and independent variables

Row	**Toyota accident**	**Toyota Crash**	**Toyota problem**
Sudden Acceleration	0.455	0.314	0.243
Acceleration Problem	0.498	0.316	0.325

In matrix form, it will be:

$$R = \begin{pmatrix} 0.455 & 0.314 & 0.243 \\ 0.498 & 0.316 & 0.325 \end{pmatrix}$$

In general, for the Radic determinant of R is given by

$$Radic\left|R_{XY}\right| = \begin{vmatrix} \rho_{x1,y1} & \rho_{x1,y2} & \rho_{x1,y3} \\ \rho_{x2,y1} & \rho_{x2,y2} & \rho_{x2,y3} \end{vmatrix} = \begin{vmatrix} \rho_{11} & \rho_{12} & \rho_{13} \\ \rho_{21} & \rho_{22} & \rho_{23} \end{vmatrix} = \begin{vmatrix} \rho_{11} & \rho_{12} \\ \rho_{21} & \rho_{22} \end{vmatrix} - \begin{vmatrix} \rho_{11} & \rho_{13} \\ \rho_{21} & \rho_{23} \end{vmatrix} + \begin{vmatrix} \rho_{12} & \rho_{13} \\ \rho_{22} & \rho_{23} \end{vmatrix}$$

In our case, $Radic\left|R_{XY}\right| = 0.025262$, which indicates a high correlation between vectors Y and X

The first step for the canonical correlation analysis (CCA) if to check the appropriateness of CCA via Rao F= 21.769 (p-Value =0.000) with global R-square = 0.317 (Jolliffe, 2002)

The canonical correlation are given by

Canonical Coefficients for Dependent (y) Set

Table 3. Coefficient for the Canonical Vectors for Dependent and Independent Variables

	$\mathbf{W_{y1}}$	$\mathbf{W_{y2}}$
Toyota accident	0.901	0.062
Toyota crash	-0.050	-0.706
Toyota problem	0.276	0.860

Canonical Coefficients for Independent (x) Set

	$\mathbf{W_{x1}}$	$\mathbf{W_{x2}}$
Sudden acceleration	0.435	-1.132
Acceleration problem	0.688	0.999

Based on the data of Table 3, as an example, the first canonical correlations are given by

$$W_{y_1} = 0.901 accident - 0.50 crash + 0.276 problem$$

$$W_{x_1} = 0.435 sudden + 0.688 acc - problem$$

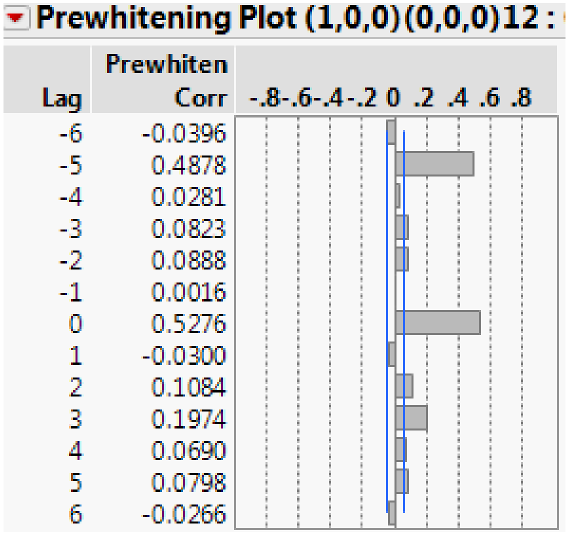
The correlation among the Canonical Correlation are given by

$$Corr(W_{1,y}, W_{1,x}) = 0.560$$

$$Corr(W_{2,y}, W_{2,x}) = 0.072$$

Let us consider the Cross Correlation analysis between the first two Canonical Correlations $W_{1,y}, W_{1,x}$ in order to develop a model such as $W_{1,y} = \beta_0 + \beta_1 W_{1,x,t-1} + \beta_2 W_{1,x,t-2} + \beta_3 W_{1,x,t-3} + ... + \beta_5 W_{1,x,t-k*} + \varepsilon_t$

The cross correlation graph after pre-whitening $W_{1,x,}$ is shown:



Prewhitening Plot (1,0,0)(0,0,0)12 :

Lag	Prewhiten Corr	-.8-.6-.4-.2 0 .2 .4 .6 .8
-6	-0.0396	
-5	0.4878	
-4	0.0281	
-3	0.0823	
-2	0.0888	
-1	0.0016	
0	0.5276	
1	-0.0300	
2	0.1084	
3	0.1974	
4	0.0690	
5	0.0798	
6	-0.0266	

Figure 4. Cross Correlation Functions for the First Canonical Correlations for the dependent and Independent variables

Namely $CCF(W_{1,t,y}, W_{1,t-5x}) = 0.4878$, which suggest the following model

$$W_{1,t}(Y) = b_0 + b_1 W_{1},_{t\text{-}5}(X)$$

and the graph showed in Figure 5.

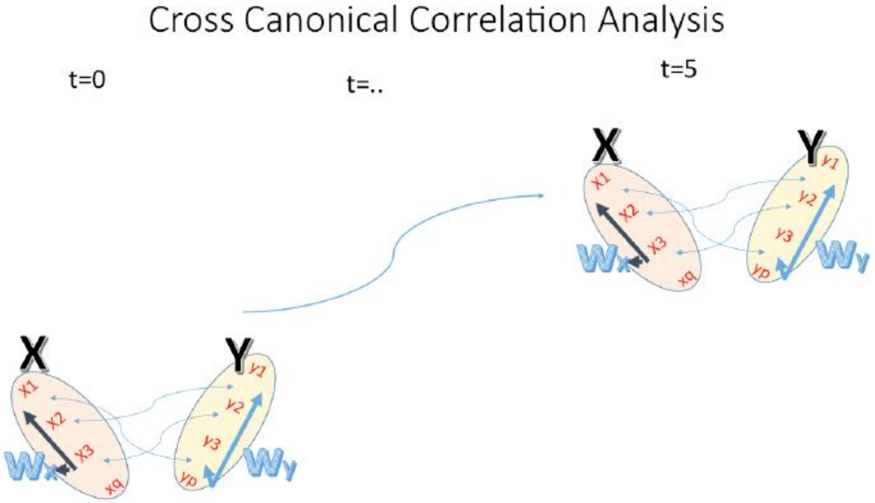


Figure 5. Graphic representation of the cross correlation between the dependent and independent variables

Conclusions and Further Research

The intention of this paper is to present an alternative approach of the VARMAX modelling taking into account the correlation within vectors Y and X, along with the inter-correlation between vectors Y and X. The canonical correlation approach reduces significantly the number of parameters to be considered and provides a useful framework to simplify the complexity of the problem. When the correlation among the component of vectors Y and X is not linear but monotonic the use of the Spearman correlation coefficient (nonparametric measure of rank correlation) is an alternative.

- The following algorithm could be of use:
    1. Select vector Y_t and X_t time series from a database such as Google Correlation
    2. Determine $|R_{XX}| \approx 1, |R_{YY}| \approx 1$; this is an indication that you can minimize the number of variables contained in vector Y and X
    3. Determine $Radic|R_{YX}| \approx 0$, namely vector Y and X are highly correlated; if this condition is not fulfilled, stop and consider using VARMAX.
    4. Apply Canonical Correlations to Model Y and X
    5. Compute and test the canonical variables for Model Y and X namely W_y and W_x
    6. Compute and test the Cross Correlation Function CCF(k^*) for W_{yt} and W_{xt}
    7. Based on 6) create a Time Series Regression Model $W_{1,y} = \beta_0 + \beta_1 W_{1,x,t-1} + \beta_2 W_{1,x,t-2} + ... + \beta_{k*} W_{1,x,t-k*} + \varepsilon_t$

- In 2) In Case of highly collinear vector Y or X , i.e., $|R_{XX}| \approx 0, or, |R_{YY}| \approx 0$, perform first Principal component Analysis.

- For the case for Non-Linear Monotonic Correlation among the variables contained in vectors Y and X, namely correlation between Y and X: Replace 1 by 1*
    1. Select vector Y_t and X_t time series from a database such as Google Correlation and replace the values of vector Y and X by their ranks.

Replacing the values of Y and X by their ranks is equivalent to use the nonparametric correlation coefficients (such as the Spearman correlation coefficient) in $|R_{XX}|, |R_{YY}|, Radic|R_{YX}|$.

References

Bilodeau, M., & Brenner, D. (1999). *Theory of multivariate statistics*. New York, NY: Spring Verlag.

Breusch, T., & Pagan, A. (1980). The lagrange lultiplier test and its applications to model specification in econometrics. *The Review of Economic Studies*, 47(1), 239-253.

Borga, M. (2001). Canonical correlation: A tutorial. Retrieved from: www.imt.liu.se/~magnus/cca/tutorial/tutorial.pdf

Chatfield, C. (2005). *Time-series forecasting*. New York, NY: Wiley Ed.

Stephens-Davidowitz, S., & Varian, H. (2015). A hands-on guide to Google data. Working paper. *Google, Inc.* Retrieved from: http://people.ischool.berkeley.edu/~hal/Papers/2015/primer.pdf

Doornik, J. (1996). *Testing vector error autocorrelation and heteroscedasticity*. Proceedings of the Econometric Society 7th World Congress, Tokyo, Japan.

Ginsberg, J. (2008). Detecting influenza epidemics using search engine query data. Retrieved from: http://research.google.com/archive/papers/detecting-influenza-epidemics.pdf

Granger, C. W. J., & Newbold, P. (1974). Spurious regression in econometrics. *Journal of Econometrics, 2*(2), 111-120.

Haig, B. (2006). Spurious correlation. *Encyclopedia of measurement and statistics*. Thousand Oaks, CA: SAGE Publications.

Jolliffe, I. (2002). *Principal component analysis (2nd ed.).* New York, NY: Springer Verlag.

Konishi, S. (1979). Asymptotic expansions for the distributions of statistics based on the sample correlation matrix in principal component analysis. *Hiroshima Math Journal, 9*(3), 647-700.

Mohebbi, M., Vanderkam, D., Kodysh, J., Schonberger, R., Choi, H., & Kumar, S. (2011). Google correlate whitepaper. Retrieved from: https://research.google.com/pubs/pub41695.html

Radic, M. (1969). A definition of the determinant of a rectangular matrix. *Glasnik Matemaicki, 1*, 1-21.

Preisendorfer, R. W., & Mobley, C. D. (1988). *Principal component analysis in meteorology and oceanography (Vol. 17).* Elsevier Science Ltd.

Quality Control Systems Corp (2011). Consumer complaints and warranty repairs in NASA-NHTSA's Toyota study of unintended acceleration, July 21, 2011. Retrieved from http://www.safetyresearch.net/Library/Report110721.pdf

Sušanj, R., & Radic, M. (1994). Geometrical meaning of one generalization of the determinant of square matrix. *Glas Mat Ser, 29*(2), 217-233.

Shumway, R. H., & Stoffer, D. S. (2010). Time series analysis and its applications: With R examples. New York, NY: Springer Science & Business Media.

Stanimirović, P., & Stanković, M. (1997). Determinants of rectangular matrices and Moore-Penrose inverse. *Novi Sad Journal of Mathematics, 27*(1), 53-69. Retrieved from

https://www.emis.de/journals/NSJOM/Papers/27_1/NSJOM_27_1_053_069.pdf

Tiao, G., & Tsay, R. (1989). Model specification in multivariate time series. *Journal of the Royal Statistical Society, Series B (Methodological), 51*(2), 157-213.

Author's Biography

Dr. Guerrero-Cusumano is a Tenured Associate Professor at the Georgetown University School of Business. He holds a Ph.D. in Industrial Engineering from the University of Illinois and a Master of Sciences in Statistics degree from the Mathematics Department of the University of Illinois and he is also an Economics Statistician from the School of Economics Science, Rosario University, Argentina. Professor Guerrero Cusumano was the former academic director of the Corporate International Master's at the Georgetown University School of Business. He was former Director of the International Institute on Government, Management, and Policy at Georgetown U. Prof. Guerrero-Cusumano has been a member of the faculty at Georgetown since 1989 in the School of Business. He is also the recipient of the Gold Medal for service at Georgetown University. In 2008, Professor Guerrero Cusumano was awarded an Honorary Doctorate, Doctor Honoris Causa in Administration by Ovidius University (Romania). Also in 2008, he has been elected a Fellow at the Judge Business School, Cambridge University, Great Britain. He was also conferred and nominated for different teaching and research awards.

Appendix: Testing Independence for Vectors Y and X

In order to test $\left|R_{XX}\right| = 1$ or $\left|R_{YY}\right| = 1$, more specifically

$H_0)$ *Uncorrelated* $\rightarrow |R| = 1$
$H_1)$ *Correlated* $\rightarrow |R| = 0$

we can use the Bartlett test, namely

$$-\left(n-1-\frac{2p+5}{6}\right)\ln|R| \overset{n\rightarrow\infty}{\approx} \chi^2_{\left(\frac{p(p-1)}{2}\right)}$$

For any alternative degree of correlatedness or independence in the normal case other than $|R| = 1$ or $|R| = 0$

$H_0)\; |R| = |P|$
$H_1)\; |R| \neq |P|$

Konishi (1979) provides the following formula:

$$\sqrt{n-1}\left(|R| - |P|\right) \overset{d}{\rightarrow} N\left\{0, 2|P|^2\left[Tr(P^2) - p\right]\right\}$$