International Journal of Mathematical, Engineering and Management Sciences
Vol. 5, No. 1, 45-55, 2020
https://doi.org/10.33889/IJMEMS.2020.5.1.004

A Comparative Study of Multivariate Analysis Techniques for Highly Correlated Variable Identification and Management

K. Ntotsis
Lab of Statistics and Data Analysis,
Department of Statistics and Actuarial-Financial Mathematics,
University of the Aegean, Greece.
E-mail: kntotsis@aegean.gr

E. N. Kalligeris
Lab of Statistics and Data Analysis,
Department of Statistics and Actuarial-Financial Mathematics,
University of the Aegean, Greece.
E-mail: ekalligeris@aegean.gr

A. Karagrigoriou
Lab of Statistics and Data Analysis,
Department of Statistics and Actuarial-Financial Mathematics,
University of the Aegean, Greece.
Corresponding author: alex.karagrigoriou@aegean.gr

(Received August 12, 2019; Accepted October 1, 2019)

Abstract
In this work we attempt is to locate and analyze via multivariate analysis techniques, highly correlated covariates (factors) which are interrelated with the Gross Domestic Product and therefore are affecting either on short-term or on long-term its shaping. For the analysis, feature selection techniques and model selection criteria are used. The case study focuses on annual data for Greece for the period 1980-2018.

Keywords- Multicollinearity, Correlation feature selection, Model selection criteria, Multivariate analysis, Principal component analysis.

1. Introduction

The purpose of this work is to identify an optimal model for the Gross Domestic Product (*GDP*). The Organization for Economic Co-operation and Development (OECD) states that "Gross Domestic Product (*GDP*) is the standard measure of the value added created through the production of goods and services in a country during a certain period. As such, it also measures the income earned from that production, or the total amount spent on final goods and services (less imports). While *GDP* is the single most important indicator to capture economic activity, it falls short of providing a suitable measure of people's material well-being for which alternative indicators may be more appropriate. This indicator is based on nominal *GDP* (also called *GDP* at current prices or *GDP* in value) and is available in different measures" (OECD, 2019). Based on well-established and proven studies it is known that GDP can be expressed by

$$GDP = C + I + G + (Ex - Im) \tag{1}$$

where *C* represents the Private Consumption Expenditures, *I* the Private Domestic Investments, *G*

International Journal of Mathematical, Engineering and Management Sciences
Vol. 5, No. 1, 45-55, 2020
https://doi.org/10.33889/IJMEMS.2020.5.1.004

the Government Consumption Expenditures, *Ex* the Total Exports and *Im* the Total Imports.

The goal of this work is to locate and analyze the interrelationships between *GDP* and various factors/variables which are interdependent and often characterized by a high degree of multicollinearity. The *GDP* is frequently used by central banks, public entities and private businesses as a standard measurement for the economic health of a country (Callen, 2008). For predictive purposes, researchers often rely on economic or financial indices and model identification procedures. den Reijer (2005) and Schumacher (2007) both studied the forecasting of Dutch and German respectively, *GDP* through factor modelling. Later, Akhter et al. (2012) used Principal Component Analysis in order to obtain a model for the *GDP* of Bangladesh. Bai et al. (2015) has shown the accuracy of factor analysis in the evaluation of the economy of a country, including variables such as Unemployment Rate, Investments, Population and General Government Total Expenditures, which are part of the current model analysis. Because of its unstable economy, Greece is the focus of many economic analyses from organizations such as OECD, Eurostat, International Monetary Fund and there is sufficient material and data on their websites one can refer to.

The explanatory variables/factors (see Table 1) that were chosen are highly correlated and result in severe multicollinearity in the primary model which appears to be a frequent problem in financial and economic big data analytics (Wang and Alexander, 2019). For the reduction or even elimination of the multicollinearity, which is a common issue in data analysis in finance and economics (Kondo et al., 2018), a number of dimension reduction techniques were used in order to identify an optimal model with a set of new uncorrelated variables/factors. In this work, for comparative purposes and for measuring the quality of each model, three information criteria were used, namely Akaike Information Criterion (Akaike, 1974), Bayesian Information Criterion (Schwarz, 1978) and Modified Divergence Information Criterion (Mantalos et al., 2010).

Table 1. Explanatory variables

Exports of Goods and Services (X_1)	Household Consumption Expenditures (X_3)	Population (X_6)
General Government Total Expenditures (X_2)	Imports of Goods and Services (X_4)	Total Labor Force (X_7)
	Investments (X_5)	Unemployment Rate (X_8)

In this work we rely on multivariate analysis and in particular, on Dimension Reduction Techniques (see e.g. Li, 2018) and Multivariate Linear Regression (see e.g. Anderson, 2003) for the modelling of the Gross Domestic Product by identifying an appropriate set of factors from a long list of possible explanatory interdependent variables which likely interact with and affect the *GDP*. The choice of *GDP* is obvious since it is a quantity of great interest for micro as well as macroeconomics. The case of Greece is chosen due to extreme economic events of recent years that greatly affected all aspects of economic activity.

The rest of the paper has been organized as follows. Section 2 provides the information and characteristics of the dataset used in this work. Section 3 discusses the Dimension Reduction Techniques methods that were used, including Principal Component Analysis (PCA) (Jolliffe,

International Journal of Mathematical, Engineering and Management Sciences
Vol. 5, No. 1, 45-55, 2020
https://doi.org/10.33889/IJMEMS.2020.5.1.004

1972; Artemiou and Li, 2009; Artemiou and Li, 2013; Wang, 2018), Beale et al. (1967) technique and Hall's (1999) Correlation-based Feature Selection (CFS) technique for the identification of the appropriate set of factors that affect the *GDP*. Section 4 deals with the modelling of *GDP* while the Assessment and Comparison is based on model identification procedures, more specifically on Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and Modified Divergence Information Criterion (MDIC). Section 5 provides the results of the optimal model. Finally, in the conclusion-discussion section, the techniques and their results are discussed along with possible extensions.

2. Preference Data

The Gross Domestic Product is interrelated, according to the relevant theory, with a variety of explanatory variables which possibly affect *GDP*.

This work is based on Greece's economy with annual data collected through Knoema, OECD and Eurostat for the eight (8) explanatory variables X_1–X_8 presented in Table 1 for the period 1980-2018 (39 annual observations). Three (3) missing values have been replaced by the average value of the preceding and the following year.

3. Dimension Reduction Techniques

3.1 Discarding Variables Technique

In order to discard variables of limited information, the Beale et al. (1967) technique was used. This technique is a simple three-step procedure proposed by Beale et al. (1967) for discarding variables in multivariate data analysis. The technique can be summed up as follows:

(i) Locate the minimum eigenvalue and the corresponding eigenvector of the variance–covariance or correlation matrix of the covariates involved.

(ii) Locate the element of the eigenvector with the highest absolute value. This value corresponds to one of the original variables (covariates) which is removed from the analysis.

(iii) Repeat steps 1 and 2 until p-k variables have been removed, where p represents the number of covariates and k represents the number of eigenvalues, which are larger than one.

The implementation of this procedure, results in the removal of Exports of Goods and Services, General Government Total Expenditures, Household Consumption Expenditures, Imports of Goods and Services, Investments and Population from the model. Thus, the technique suggests the use of the Total Labor Force and the Unemployment Rate as the only variables interrelated with *GDP* and thus affecting its modelling.

3.2 Principal Component Analysis

We proceed now with the implementation of the Principal Component Analysis as an alternative dimension reduction technique and manage to obtain the complete set of the 8 principal components, with the corresponding eigenvalues ranging from around 6.5 to nearly zero. This technique was proposed independently by Pearson (1901) and Hotelling (1933; 1936). The idea behind PCA is the conversion of a dataset with interdependent variables into a new one with uncorrelated variables (principal components), which are arranged in such a way so that the first few components maintain the greater part of the variability that exists among the original variables. Under this procedure the dimensionality reduction of the original data set can be achieved while

International Journal of Mathematical, Engineering and Management Sciences
Vol. 5, No. 1, 45-55, 2020
https://doi.org/10.33889/IJMEMS.2020.5.1.004

leaving unchanged as much as possible the variation (Jolliffe, 2002).

The components constitute a set of uncorrelated vectors which have been created by the following methodology:

Let us denote by C_j the j^{th} component, λ_j the corresponding eigenvalue and $v_{ij} = (v_{1j}, \dots, v_{nj})'$ the corresponding eigenvector, $i = 1, \dots, n,\ j = 1, \dots, m$ where n is the number of observations and m is the total number of original variables/covariates. Hence, C_j is defined as:

$$\mathrm{C_j} = -\sqrt{\lambda_j} * v_{ij} = -\sqrt{\lambda_j} * \begin{pmatrix} v_{1j} \\ v_{2j} \\ \vdots \\ v_{nj} \end{pmatrix} \tag{2}$$

$$\forall\, i = 1, \dots, n, \qquad j = 1, \dots, m, \qquad n = 39, \qquad m = 8.$$

Each new component/variable $Z_j,\ j = 1, \dots, m$ is a linear combination of the component's matrix and the original dataset matrix. Indeed, let us denote by

X_{ij}: the elements of the original dataset matrix,
C_{jj}: the elements of the components matrix and
Z_{ij}: the elements of the new variables matrix.

Then, the Z_{ij} element of the vector $Z_j = (Z_{1j}, \dots, Z_{8j})'$ is:

$$Z_{ij} = \sum_{j=1}^{m} X_{ij} C_{ij},\ \forall\, i = 1, \dots, n,\ j = 1, \dots, m \tag{3}$$

Based on the overall results and the fact that it is preferable to avoid the loss of important information, we conclude that the first two components (Z_1 and Z_2) should be kept (see Table 2) regardless of the eigenvalues because they retain a considerable amount of the total information/variability (more than 95% of the original variability of the data). The described variability played a key role in the aforementioned decision since the intention was to keep that many components so that a considerable proportion of the original variability will be described by the components chosen.

Table 2. The Two primary principal components

1st Component (Z1)		2nd Component (Z2)
General Government Total Expenditures (0.97)	Imports of Goods and Services (0.97)	Investments (0.62)
Household Consumption Expenditures (0.99)	Total Labor Force (0.97)	Unemployment Rate (0.74)

International Journal of Mathematical, Engineering and Management Sciences
Vol. 5, No. 1, 45-55, 2020
https://doi.org/10.33889/IJMEMS.2020.5.1.004

Remark: To determine which variables are significant in each component, the following empirical rule was followed. For the two chosen components, the variables for which the absolute value of the associated coefficient is at least equal to 0.95 are kept as significant. A value of around 0.95, although there is no specific rule, is considered to be satisfactory in retaining a sufficient amount of information.

For the problem at hand, the first component, denoted by Z_1, holds more than 80% of the total variation of the dataset while the second one, denoted by Z_2, holds roughly 15% of it. The rest of the components contain the remaining percentage of variation. By construction, the first component is considered to be the most important in which the analysis is primarily based on. Having said that, we observed in the above analysis, 6 of the total of 8 variables emerge as important according to the associated coefficients given in parenthesis (see Table 2).

Remark: For modelling purposes both PCA significant variables/components (Z_1 and Z_2) are used in their full form that contains, not only the significant variables (with coefficients at least equal to 0.95) which are presented in Table 2, but all $m=8$ original X_i's in (3).

As it can be seen from Table 2, General Government Total Expenditures, Household Consumption Expenditures, Imports of Goods and Services and Total Labor Force emerge as important in the first component while Investments and Unemployment Rate in the second one.

Hence, using this technique we proceed with the Multivariate Analysis of the Gross Domestic Product with Z_1 and Z_2 as the uncorrelated variables affecting *GDP*.

3.3 Correlation-Based Feature Selection

Variable selection, also known as feature selection (Guyon and Elisseeff, 2003), is the procedure of evaluating all possible subsets of a dataset and finding the one that minimizes the error rate. Through this process, the best subset of relevant variables will emerge for a better model construction. Furthermore, all insignificant variables will be removed without incurring much loss of information. In this work a Correlation-based Feature Selection, denoted by CFS (Hall, 1999), will be used as the third approach for Dimension Reduction. CFS is a measure that evaluates subsets of features on the basis of the following Hall's hypothesis:

An optimal feature subset includes uncorrelated independent covariates (features) and simultaneously high correlations between each covariate with the dependent variable. If such correlations are available, then the merit of a feature subset S consisting of N features is defined as:

$$Merit_{S_N} = \frac{N\,\overline{r_{YX_i}}}{\sqrt{N+N(N-1)\overline{r_{X_iX_i}}}} \tag{4}$$

where $Merit_{S_N}$ is the correlation between the summed independent variables and the dependent variable, N is the number of variables, $\overline{r_{YX_i}}$ is the average of the correlations between the independent variables and the dependent variable, and $\overline{r_{X_iX_i}}$ is the average inter-correlation between the independent variables. Hall presented a backward elimination procedure, with the use of (4) in order to choose a subset. The full set of variables is evaluated with (4), which, in fact, is the Pearson's correlation coefficient with standardized variables. Then, a variable is temporarily removed and the set of variables is evaluated with the aforementioned equation. If the subset scores

International Journal of Mathematical, Engineering and Management Sciences
Vol. 5, No. 1, 45-55, 2020
https://doi.org/10.33889/IJMEMS.2020.5.1.004

are higher than the set before, then the variable is permanently removed. Otherwise, it is reinstated. The process continues until each variable is removed once and the effect of its removal is measured. The process stops when no subset scores are higher than those of the original set.

The implementation of this procedure in the examined dataset, results in the withdrawal of 5 out of the total 8 original variables. The remaining variables, namely General Government Total Expenditures, Household Consumption Expenditures and Imports of Goods and Services are considered as the important ones in the modelling of *GDP*. It must be noted that the same variables together with the Total Labor Force compose the first and most important component (Z_1), of PCA.

3.4 Techniques Review

In the previous sections three-dimension reduction techniques were implemented for the identification of interrelationships between a number of potentially significant factors and the *GDP*. While in some cases similarities between the techniques were revealed, all three highlight different variables as important, as it can be seen in Table 3.

Table 3. Dimension reduction techniques synopsis

Beale et al. (1967)	PCA		CFS
Total Labor Force	General Government Total Expenditures	Investments	General Government Total Expenditures
Unemployment Rate	Household Consumption Expenditures	Total Labor Force	Household Consumption Expenditures
	Imports of Goods and Services	Unemployment Rate	Imports of Goods and Services

4. Model Selection Criteria

Model identification procedures play a pivotal role in statistics by identifying the best model among an available class of models. Those techniques are contemplated as assessors of a quantity. For example, for a given data the probability of the proposed model can be used as assessor which is essential for the pursuit of identifying the optimal fundamental structure of the phenomenon under investigation.

Model identification procedures have been heuristically recommended for time varying processes. Kullback and Leibler (1951) developed such a measure that minimizes the loss of information. A direct connection between the Kullback-Leibler (KL) measure and the maximum likelihood estimation (MLE) method, gave rise to the well-known Akaike Information Criterion (AIC, Akaike, 1974). A related procedure also associated with the likelihood function is the Bayesian Information Criterion (BIC, Schwartz, 1978). These criteria are the most popular ones, among others. In this work, in addition to AIC and BIC, a recently developed information criterion known as Modified Divergence Information Criterion (MDIC), proposed by Mantalos et al. (2010), will be used for comparative purposes.

4.1 Akaike Information Criterion

The AIC can be considered as the relative amount of information lost by the candidate model: the less information lost, the higher the model's quality. In other words, AIC approximates the quality

International Journal of Mathematical, Engineering and Management Sciences
Vol. 5, No. 1, 45-55, 2020
https://doi.org/10.33889/IJMEMS.2020.5.1.004

of a candidate model relative to each of the other candidate models for the data. As mentioned above, the task is accomplished by combining a criterion that minimizes the loss of information with a maximum likelihood estimation method. More specifically, AIC is based on the log-likelihood function and is defined as:

$$AIC_p = -2(maximum\ log-likelihood) + 2p \tag{5}$$

where *p* represents the dimension of the vector-parameter θ. The optimal model is the one with the lowest *AIC* value.

4.2 Model Selection based on AIC

In the previous section three-dimension reduction/variable selection techniques were used in order to find the optimal explanatory variables for the modelling of Gross Domestic Product, namely, Beale et al. (1967), PCA and Hall's CFS Selection technique. Using *GDP* as the dependent variable and the selected variables of each technique as the independent ones, the following three models were constructed corresponding to Beale et al. (1967), PCA and CFS techniques respectively:

$$Y_i = \alpha_{11} + \beta_{11}X_{i7} + \beta_{12}X_{i8} + \varepsilon_{i1}\ , \qquad i = 1, \dots, 39$$

$$Y_i = \alpha_{21} + \beta_{21}Z_{i1} + \beta_{22}Z_{i2} + \varepsilon_{i2}\ , \qquad i = 1, \dots, 39$$

$$Y_i = \alpha_{31} + \beta_{31}X_{i2} + \beta_{32}X_{i3} + \beta_{33}X_{i4} + \varepsilon_{i3}\ , \qquad i = 1, \dots, 39.$$

From the results in Table 4, it appears that the optimal model based on Akaike Information Criterion is the one formulated by Hall's Correlation-based Feature Selection technique and contains the General Government Total Expenditures, the Household Consumption Expenditures and the Imports of Goods and Services as the independent variables.

4.3 Bayesian Information Criterion

The Bayesian information criterion is a model identification procedure based on information theory but set within a Bayesian context. It is an evaluation criterion for models estimated by using the maximum likelihood method. *BIC* can be considered as an estimate of a function of the posterior probability of a model being true, under a certain Bayesian setup, so that a lower *BIC* means that a model is considered to be more likely to be the true model. BIC is given by

$$BIC_P = -2(maximum\ log-likelihood) + p\log n \tag{6}$$

where *p* represents the dimension of the vector-parameter θ and *n* is the number of observations.

Laplace method for integrals has been used for obtaining the marginal likelihood associated with the posterior probability in (6). The results of the implementation of BIC can be seen in Table 4, where we observe that *BIC*, like *AIC* chooses Hall's CFS as the best model.

4.4 Modified Divergence Information Criterion

The Divergence Information Criterion (*DIC*) proposed by Mattheou et al. (2009) constitutes a modelling generalization of *AIC*, based on the Basu, Harris, Hjort, and Jones (BHHJ) divergence measure (Basu et al., 1988). *DIC* family of procedures, like *AIC*, is an asymptotic approximation as the sample size increases and offers an alternative based on the so called divergence measures.

International Journal of Mathematical, Engineering and Management Sciences
Vol. 5, No. 1, 45-55, 2020
https://doi.org/10.33889/IJMEMS.2020.5.1.004

In this work, we consider the Modified Divergence Information Criterion (*MDIC*), which consists a modification of *DIC*, proposed by Mantalos et al. (2010). *MDIC* can be viewed as an approximation of the expected overall discrepancy, which based on the BHHJ measure, evaluates the distance between the true and the fitted models. If the model with the smallest estimator of the expected overall discrepancy is chosen, then it is possible to end up with a model with an unnecessarily large number of variables. Thus, the Modified Divergence Information Criterion is a criterion comparable to *AIC*. *MDIC* is given by

$$MDIC_p = n\, MQ_{\hat{\theta}} + (2\pi)^{-\frac{\alpha}{2}} (1+\alpha)^{2+\frac{p}{2}} p \tag{7}$$

where

- p is the order of the model or the number of variables involved.
- $MQ_{\hat{\theta}} = -\left[(1+\frac{1}{\alpha})\frac{1}{n}\sum_{i=1}^{n} f_{\hat{\theta}}^{\alpha}(x_i)\right]$.
- $\hat{\theta}$ is a consistent and asymptotically normal estimator of the parameter vector θ.
- $\alpha \in (0,1)$ is the positive index, often chosen to be equal to 0.25 (see Mantalos et al., 2010).

The results based on *MDIC* together with those based on *AIC* and *BIC* are provided in Table 4.

Table 4. Model selection summary

	AIC	***BIC***	***MDIC***
Beale et al. (1967)	2009.233	2015.888	5.206
PCA	1925.267	1941.867	37.896
CFS	1901.246	1909.564	7.761

5. Conclusion and Future Research

In conclusion, in this paper, we attempted via dimension reduction techniques, to identify interrelationships between the Gross Domestic Product of Greece and a number of factors which are highly correlated. Beale et al. (1967), Principal Component Analysis and Hall's Correlation-based Feature Selection techniques were implemented and suggested different models with different variables (see Table 3).

More specifically, Beale et al. (1967) proposed a model with the Total Labor Force (X_7) and the Unemployment Rate (X_8) as independent variables. This technique clearly focuses solely on the workforce point of view in order to achieve the optimal model. PCA, on the other hand, instead of using the original variables, created new uncorrelated ones. In fact, PCA promotes a model with two *uncorrelated* variables (Z_1 and Z_2). Through them, 6 out of a total of 8 variables emerge as important, namely X_2, X_3, X_4, X_5, X_7 and X_8 (see Table 2). It should be noted that the variables selected as significant have also been chosen either by Beale's or Hall's models. The third technique, CFS, proposed a model with the General Government Total Expenditures (X_2), the Household Consumption Expenditures (X_3) and the Imports of Goods and Services (X_4) as significant variables affecting *GDP*.

Based on theoretical background (see Equation (1)), it appears that the CFS model covers most part of *GDP*'s formula and seems to be able to identify and select the" right" subset of variables from the original ones. Indeed, although CFS does not select the Investments and the Exports of Goods and Services which both are part of the variables involved in (1), it is able to identify, the Imports of Goods and Services (which is part of the Imports), the Government Expenditures and the Household Consumption Expenditures. Note though that CFS model also chooses to ignore demographic variables, which affect indirectly and not directly the modelling of *GDP* through their interrelationships with all variables involved in (1).

The theoretical interpretation of the results is confirmed by two out of the three model selection criteria that were used and their results are provided in Table 4. Both *AIC* and *BIC* select Hall's CFS model, while MDIC selects Beale et al. (1967) model.

From the analysis, we see that the PCA model is not the optimal in all cases examined. When it comes to CFS and Beale et al. (1967), we observe that, both *AIC* and *BIC*, choose clearly the former, leaving way behind the latter. On the other hand, although MDIC is in favor of Beale et al. (1967), the difference observed as compared to CFS, could not be considered significant.

The main obstacle that we had to overcome in this work was the problem of multicollinearity, which is very common especially when it comes to modelling that involves big data on various financial characteristics and/or economic indicators. The case of the *GDP* of Greece was an ideal example to explore the capabilities of various multivariate analysis techniques in handling the multicollinearity problem and identify a set of influential factors. Taking that under consideration, it is possible, in a future work, to attempt to explore how different model selection criteria react or are able to make the right variable/model selection, when multicollinearity is of different magnitude (non-existent, low, medium, high, nearly perfect or even perfect). Through this process one could be able to identify the criterion which is better adjusted and finally succeeds in choosing the optimal model when the variables involved are highly correlated.

Conflict of Interest
The authors declare that there is no conflict of interest in the article contents.

Acknowledgements
The authors wish to express their appreciation to the Editor and an anonymous referee whose comments and suggestions improved both the quality and the presentation of the manuscript. This work was completed as part of the research activities of the *Laboratory of Statistics and Data Analysis* of the University of the Aegean.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716-723.

Akhter, Y., Mahsin, M.D., & Mohaimin, M.Z. (2012). An application of factor analysis on gross domestic product data of Bangladesh. *Bangladesh e-Journal of Sociology*, *9*(1), 6-18.

Anderson, T.W. (2003). *An introduction to multivariate statistical analysis*. New York, Wiley.

Artemiou, A., & Li, B. (2009). On principal components and regression: a statistical explanation of a natural phenomenon. *Statistica Sinica*, *19*(4), 1557-1565.

Artemiou, A., & Li, B. (2013). Predictive power of principal components for single-index model and sufficient dimension reduction. *Journal of Multivariate Analysis*, *119*, 176-184.

Bai, A., Hira, S., & Deshpande, P.S. (2015). An application of factor analysis in the evaluation of country economic rank. *Procedia Computer Science*, *54*, 311-317.

Basu, A., Harris, I.R., Hjort, N.L., & Jones, M.C. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika*, *85*(3), 549-559.

Beale, E.M.L., Kendall, M.G., & Mann, D.W. (1967). The discarding of variables in multivariate analysis. *Biometrika*, *54*(3-4), 357-366.

Callen, T. (2008). What is gross domestic product? *Finance and Development*, *45*(4), 48-49.

den Reijer, A.H. (2005). *Forecasting Dutch GDP using large scale factor models*. DNB Working Papers 028, Netherlands Central Bank, Research Department.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*(Mar), 1157-1182.

Hall, M.A. (1999). *Correlation-based feature selection for machine learning*. The University of Waikato, Hamilton, New Zealand.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*(6), 417-441.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, *28*(3-4), 321-377.

Jolliffe, I.T. (1972). Discarding variables in a principal component analysis. I: Artificial data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *21*(2), 160-173.

Jolliffe, I.T. (2002). *Principal components analysis. 2nd Ed*., Springer-Verlag, New York.

Kondo, M., Mizuno, O., & Choi, E.H. (2018). Causal-effect analysis using Bayesian LiNGAM comparing with correlation analysis in function point metrics and effort. *International Journal of Mathematical, Engineering and Management Sciences*, *3*(2), 90-112.

Kullback, S., & Leibler, R.A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79-86.

Li, B. (2018). *Sufficient dimension reduction: methods and applications with R*. Chapman and Hall/CRC. New York.

Mantalos, P., Mattheou, K., & Karagrigoriou, A. (2010). An improved divergence information criterion for the determination of the order of an AR process. *Communications in Statistics—Simulation and Computation*, *39*(5), 865-879.

Mattheou, K., Lee, S., & Karagrigoriou, A. (2009). A model selection criterion based on the BHHJ measure of divergence. *Journal of Statistical Planning and Inference*, *139*(2), 228-235.

OECD (2019). *Organisation for economic co-operation and development definition for gross domestic product*, https://data.oecd.org/gdp/gross-domestic-product-gdp.htm (as of Aug. 4, 2019).

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*(11), 559-572.

Schumacher, C. (2007). Forecasting German GDP using alternative factor models based on large datasets. *Journal of Forecasting*, *26*(4), 271-302.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461-464.

Wang, L. (2018). Big data and IT network data visualization. *International Journal of Mathematical, Engineering and Management Sciences*, *3*(1), 9-16.

Wang, L., & Alexander, C.A. (2019). Big data analytics in healthcare systems. *International Journal of Mathematical, Engineering and Management Sciences*, *4*(1), 17-26.