International Journal of Theoretical and Applied Mathematics
2019; 5(2): 21-30
http://www.sciencepublishinggroup.com/j/ijtam
doi: 10.11648/j.ijtam.20190502.11
ISSN: 2575-5072 (Print); ISSN: 2575-5080 (Online)

Review Article

A Review of Constrained Principal Component Analysis (CPCA) with Application on Bootstrap

Alaa Ahmed Abd Elmegaly

Department of Advanced Management Sciences, Higher Institute of Advanced Management Sciences and Computers, Al-Buhayrah, Egypt

Email address:
bintmasr880@yahoo.com

To cite this article:
Alaa Ahmed Abd Elmegaly. A Review of Constrained Principal Component Analysis (CPCA) with Application on Bootstrap. *International Journal of Theoretical and Applied Mathematics*. Vol. 5, No. 2, 2019, pp. 21-30. doi: 10.11648/j.ijtam.20190502.11

Received: August 10, 2019; **Accepted**: August 26, 2019; **Published**: September 10, 2019

Abstract: Linear model (LM) provide the advance in regression analysis, where it was considered an important statistical development of the last fifty years, following general linear model (GLM), principal component analysis (PCA) and constrained principal component analysis (CPCA) in the last thirty years. This paper introduce a series of papers prepared within the framework of an international workshop. Firstly, the LM and GLM has been discussed. Next, an overview of PCA has been presented. Then constrained principal component has been shown. Some of its special cases such as PCA, Canonical correlation analysis (CANO), Redundancy analysis (RA), Correspondence analysis (CA), Growth curve models (GCM), Extended growth curve models (ExGCM), Canonical discriminant analysis (CDA), Constrained correspondence analysis, non-symmetric correspondence analysis, Multiple Set CANO, Multiple Correspondence Analysis, Vector Preference Models, Seemingly unrelated regression (SUR), Weighted low rank approximations, Two-Way canonical decomposition with linear constraints, and Multilevel RA has been noted in this paper. Related methods and ordinary least squares (OLS) estimator as a special case form CPCA has been introduced. Finally, an example has been introduced to indicate the importance of CPCA and the different between PCA and CPCA. Where CPCA is a method for structural analysis of multivariate data that combine features of regression analysis and principal component analysis. In this method, the original data first decomposed into several components according to external information. The components then subjected to principal component analysis to explore structures within the components.

Keywords: General Linear Model, Principal Component Analysis, Constrained Principal Component Analysis, Bootstrap

1. Introduction

LM play a central part in modern statistical methods these models are able to approximate a large amount of metric data structures in their entire range of definition or at least piecewise. On the other hand, approaches such as the analysis of variance, which model effects such as linear deviations from a total mean, have proved their flexibility, and error structures of most ecological data.

According to Gauss Markov theorem, which is based on the linear regression model (LM),

$$Y_{n.1} = X_{n.p}\beta_{p.1} + \epsilon_{n.1} \tag{1}$$

where Y is an $n \times 1$ vector of responses, X is an $n \times p$ observed matrix of the variables, assumed to have full rank, i.e., rank $(X) = p$, β is a $p \times 1$ vector of unknown parameters, and ϵ is an $n \times 1$ vector of error terms assumed to be multivariate normally distributed with mean 0 and variance covariance $\sigma^2 I_p$. It is known that the ordinary least squares (OLS) estimator of β

$$\hat{\beta}_{OLS} = (\acute{X}X)^{-1}\acute{X}Y, \tag{2}$$

which distributed normal $\mathcal{N}(\beta, \sigma^2(\acute{X}X)^{-1})$. The standard regression model assumes that the column vectors in X are linearly independent. The restricted model for $\hat{\beta}_{OLS}$ can be written as $r = R\beta$ where R is an q x p matrix$(q \leq p)$, and r is q x 1 vector of restrictions, the restricted parameter β^{C}_{OLS} using Lagrange function is given by

$$L = (Y - X\beta)'(Y - X\beta) + \lambda(r - R\beta_{OLS})$$

$$\frac{\partial}{\partial\beta}L = -2\acute{X}Y + 2(\acute{X}X)\,\beta_{OLS}^{C} + \acute{R}\,\lambda = 0$$

$$\frac{\partial}{\partial\lambda}L = r - R\beta_{OLS}^{C} = 0$$

$$\lambda = -2(R(\acute{X}X)^{-1}\acute{R}\,)^{-1}(r - R\hat{\beta}_{OLS})$$

$$\beta_{OLS}^{C} = \hat{\beta}_{OLS} + (\acute{X}X)^{-1}\acute{R}\,(R(\acute{X}X)^{-1}\acute{R}\,)^{-1}(r - R\hat{\beta}_{OLS}) \quad (3)$$

GLM is mathematical extension of linear models that do not force data into unnatural scales allow for non-linearity and non-constant variance structures in the data. They are based on an assumed relationship (called a link function) between the mean of the response variable and the linear combination of the explanatory variables. Data may be assumed to be from several families of probability distributions, including the normal, binomial, Poisson, negative binomial, or gamma distribution, many of which better fit the non-normal Thus, GLM are more flexible and better suited for analyzing ecological relationships, which can be poorly represented by classical Gaussian distributions [2].

The next sections indicate the principal component analysis PCA and the constrained principal component analysis CPCA has been shown in the third section, where some special and related cases has been introduced in the fourth section, the fifth section proves that the constrained OLS estimator is a special case from CPCA, where example on PCA and CPCA has been introduced in the sixth section, but the seventh section shows another way to analyze the data, the last section uses bootstrap to study different sample size n.

2. Principal Component Analysis

PCA was introduced in 1901 [12], it is a multivariate technique that analyzes a data in which observations are described by several inter correlated quantitative dependent variables. Its goal is to get the important information from the data, to represent it as a set of new orthogonal (independent) variables called principal components. Mathematically, PCA depends on the eigen decomposition of positive semi definite matrices and on the singular value decomposition SVD of rectangular matrices [7] and In case of multicollinearity problem, the researchers used another forms to estimate the parameters like principal component regression PCR [1]. Where this problem occurs when the predictors included in the linear model are highly correlated with each other. When this is the case, the matrix $\acute{X}\,X$ tends to be singular and hence identifying the least squares estimates will face numerical problems. Researchers used the orthogonal matrix T in the GLM to get the PCR estimator for β [3, 9, 10]:

$$Y_{n.1} = X_{n.p}T_{P.P}\,\acute{T}_{P.P}\,\beta_{p.1} + \epsilon_{n.1} \quad (4)$$

They made spectral decomposition of the matrix $\acute{X}$X given as

$$\acute{X}X = \left(T_r, T_{p-r}\right)\begin{pmatrix}\Lambda_r & 0\\ 0 & \Lambda_{p-r}\end{pmatrix}\begin{pmatrix}T_r\\ T_{p-r}\end{pmatrix} \quad (5)$$

Where $\Lambda_{\mathrm{r}} = \acute{\mathrm{T}}_{\mathrm{r}}\acute{X}\mathrm{X}\mathrm{T}_{\mathrm{r}}$ is diagonal matrix such that the main diagonal elements are the r largest eigenvalues of $\acute{X}$X, while the main diagonal elements of the $\Lambda_{\mathrm{p-r}}$ matrix are the remaining $\mathrm{p}-\mathrm{r}$ eigenvalues.

The PCR estimator for β can be written as

$$\hat{\beta}_{PC} = T_r\,(\acute{\mathrm{T}}_{\mathrm{r}}\acute{X}\,XT_r)^{-1}\,\acute{\mathrm{T}}_{\mathrm{r}}\acute{X}\,Y \quad (6)$$

Expectation and variance:

$$\mathbb{E}(\hat{\beta}_{PC}) = T_r\acute{\mathrm{T}}_{\mathrm{r}}\beta = I_r\beta\ (biased)$$

$$var(\hat{\beta}_{PC}) = \sigma^2\,T_r\,(\Lambda_{\mathrm{r}})^{-1}\acute{\mathrm{T}}_{\mathrm{r}}$$

3. Constrained Principal Component Analysis

It is a method for structural analysis of multivariate data that combines features of regression analysis and principal component analysis. In this method, the original data are first decomposed into several components according to external information. The components are then subjected to principal component analysis to explore structures within the components [17].

The constrained principal component model is:

$$Z_{N.n} = G_{N.p}M_{p.q}\acute{H_{q.n}} + B_{N.q}\acute{H_{q.n}} + G_{N.p}C_{p.n} + E_{N.n} \quad (7)$$

where Z is an $N \times n$ matrix of responses, $G\ and\ H$ are observed matrices of the variables, assumed to have full rank, $M, B, and\ C$ are matrices of unknown parameters, and E is an $N \times n$ matrix of error terms assumed to be multivariate normally distributed with mean 0 and variance covariance $\sigma^2 I$. Statistical researchers estimated the unknown matrices of parameter as [14]:

$$\widehat{M} = (\acute{G}KG)^{-}\acute{G}KZLH(\acute{H}\,LH)^{-} \quad (8)$$

$$\hat{B} = K^{-}KQ_{G/K}ZLH(\acute{H}LH)^{-} \quad (9)$$

$$\hat{C} = (\acute{G}\,KG)^{-}\acute{G}\,KZ\acute{Q_{H/L}}LL^{-} \quad (10)$$

$$\hat{E} = P_{G/K}Z\acute{P}_{H/L} - K^{-}KQ_{G/K}Z\acute{P}_{H/L} - P_{G/K}Z\acute{Q_{H/L}}LL^{-}$$

Where:

$$Q_{G/K} = I - P_{G/K},\ P_{G/K} = G(\acute{G}\,KG)^{-}\acute{G}\,K$$

$$Q_{H/L} = I - P_{H/L},\ P_{H/L} = H(\acute{H}LH)^{-}\acute{H}L,$$

K, a symmetric nnd (nonnegative definite) matrix of order N denote the cases Metric matrix, and L, a symmetric nnd (nonnegative definite) matrix of order n, to denote the variables metric matrix. If K and/or L are psd (positive-semidefinite) but not pd (positive definite), the conditions: rank (K G) = rank (G), and rank (L H) = rank (H) has been required, These conditions are essential for projectors [14], When $K = I$ and $L = I$. Putting the estimates of $M, B, C, and\ E$ above in model (7) yields the following decomposition of the data matrix:

$$Z = P_{G/K}Z\acute{P}_{H/L} + K^{-}KQ_{G/K}Z\acute{P}_{H/L} + P_{G/K}Z\acute{Q}_{H/L}LL^{-} + (Z - P_{G/K}Z\acute{P}_{H/L} - K^{-}KQ_{G/K}Z\acute{P}_{H/L} - P_{G/K}Z\acute{Q}_{H/L}LL^{-}) \quad (11)$$

3.1. Kinds of External Information

There are two kinds of matrices of external information, one on the cases and the other on the variables side of the data matrix. The former by an N x p matrix G, and the latter by an n x q matrix H has been denoted. When there is no special case and/or variable information, $G = I_N$ and/or $H = I_n$ may be set, When the rows of a data matrix represent cases, cases demographic information, such as IQ, age, and level of education, etc., may be used in G. For example, a matrix of dummy variables for G indicating cases' group membership may be taking, then analyze the differences among the groups, When the columns of a data matrix represent stimuli, a matrix of descriptor variables of the stimuli as H may be taking. When the columns correspond to different within subject experimental conditions, H could be a matrix of contrasts, or when the variables represent repeated observations, H could be a matrix of trend coefficients. There are several potential advantages of incorporating external information [16]. The empirical validity of hypotheses incorporated as external constraints by evaluating the goodness of fit of the hypotheses may be investigated, missing values via external constraints which serve as predictor variables may be predicted. In some cases incidental parameters by reparameterizing them as linear combinations of a small number of external constraints may be eliminated [13].

3.2. Internal Analysis

In the internal analysis, the decomposed matrices in (11) are subject to PCA either separately or some of the terms combined. Decisions as to which term or terms are subjected to PCA, and which terms are to be combined, are dictated by researchers' own empirical interests. For example, PCA of the first term in (11) reveals the most prevailing tendency in the data that can be explained by both G and H, while that of the fourth term is meaningful as a residual analysis [13].

4. Some Special Cases and Related Methods of CPCA

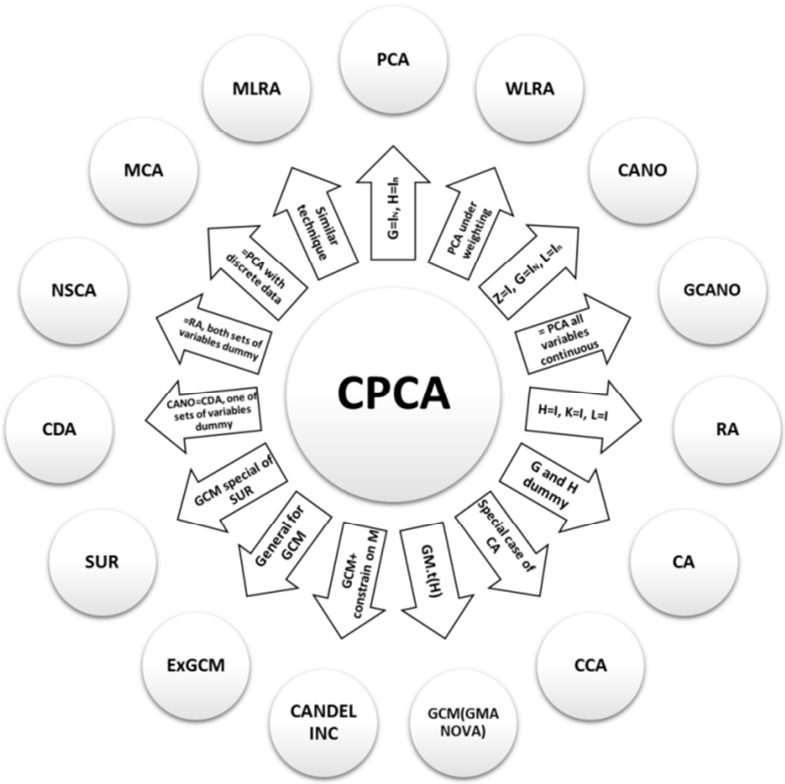

Figure 1. *Some Special cases and related methods of CPCA.*

Statistical researchers introduced about 20 special cases and related techniques for CPCA as PCA [14], CANO, and Redundancy analysis (RA) the next part indicates some of them and illustrate the assumptions that lead each case to CPCA.

1. CPCA reduces to unconstrained PCA when there is no additional case or variable information to be incorporated in the analysis. In this case $G = I_N$ and $H = I_n$ can be set, researcher also usually assume that $K = I_N$ and $L = I_n$ [15].
2. Canonical correlation analysis CANO analyzes relationships between two sets of variables. CANO can be derived from CPCA in two different ways. One is by setting $Z = I, K = I, and\ L = I$. The other is by setting $Z = (\acute{G}\,G)^{-}\acute{G}\,H(\acute{H}\,H)^{-}, K = \acute{G}\,G, L = \acute{H}\,H, G = I, \text{and } H = I$ [5].
3. RA is a useful technique for multivariate predictions. It extracts a series of orthogonal components from predictor

variables that successively account for the maximum variability in criterion variables. It maximizes the proportion of the total sum of squares in the criterion variables that can be accounted for by each successive component. The set of components thus obtained defines, in the space of the predictor variables, a subspace best predictive of the criterion variables. This is in contrast with canonical correlation analysis CANO between two sets of variables, in which components are extracted from each set that are maximally correlated with each other. A large canonical correlation, however, does not imply that the two sets of variables are highly correlated as a whole [8]. RA follows from CPCA by setting $H = I, K = I$ *and* $L = I$.

4. Correspondence Analysis (CA) When both G and H consist of dummy coded categorical variables, CANO specializes in correspondence analysis CA of a probability table $F = \acute{G}H$.
5. Multidimensional Scaling MDS, In MDS we represent both rows (cases) and columns (variables) of a data matrix in a multidimensional Euclidean space in such a way that those variables chosen by particular cases are located close to the subjects, while those variables not chosen by those cases are located far from them [14].
6. Growth Curve Models GCM also known as GMANOVA (generalized multivariate analysis of variance), provide useful methods for analyzing patterns of change in repeated measurements, and investigating how such patterns are related to various characteristics of cases.
7. Extended Growth Curve Models ExGCM, it is a generalization of GCM which has more than one structural term like $GM\acute{H}$ (the first term in the CPCA).
8. Canonical Discriminant Analysis CDA, when one of two sets of variables in CANO consists of dummy coded categorical variables, CANO reduces to canonical discriminant analysis CDA.
9. Constrained correspondence analysis, Nonsymmetric correspondence analysis, Multiple Set CANO, Multiple Correspondence Analysis, Vector Preference Models, Seemingly Unrelated Regression (SUR), Weighted Low Rank Approximations, Two-Way canonical decomposition with linear constraints, and Multilevel RA are also special cases of CPCA [14]. Researchers said that the constrained principal component model is a general model for any constrained estimator; the following section show that the constrained ordinary least square is a special case from CPCA [14].

5. The OLS Estimator Is a Special Case Form CPCA

$$\hat{Z}_{N.n} = G_{N.p}\hat{M}_{p.q}\acute{H}_{q.n} + \hat{B}_{N.q}\acute{H}_{q.n} + G_{N.p}\hat{C}_{p.n}$$
$$= G\,(\acute{G}\,KG)^{-}\acute{G}\,KZLH(\acute{H}\,LH)^{-}\acute{H} + K^{-}KQ_{G/K}ZLH(\acute{H}\,LH)^{-}\acute{H} + G\,(\acute{G}\,KG)^{-}\acute{G}\,KZQ_{H/L}^{\acute{}}LL^{-}$$
$$= G\,(\acute{G}\,KG)^{-}\acute{G}\,KZLH(\acute{H}\,LH)^{-}\acute{H} + K^{-}K(I - G(\acute{G}\,KG)^{-}\acute{G}\,K)ZLH(\acute{H}\,LH)^{-}\acute{H} + G\,(\acute{G}\,KG)^{-}\acute{G}\,KZ(I - \acute{L}\,H(\acute{H}\,LH)^{-}\acute{H}\,)LL^{-}$$

Let: The matrices $(\acute{G}\,KG)$, and $(\acute{H}\,LH)$ are non-singular, $Z = G_{N.P}S_{P.n}$, $G = X$, $K_{N.N} = I_N$, $L_{n.n} = \acute{L} = (\acute{G}G)^{-1}$ This is means that $n = p$, $H = \acute{R}$, $R\acute{S} = r$, and R Full column rank [11]

Where S is a p x n matrix and: $S = (M_{p.q}\acute{H}_{q.n} + C_{p.n})$

$$\hat{Z}_{N.n} = G\,(\acute{G}\,G)^{-}\acute{G}\,Z[LH(\acute{H}\,LH)^{-}\acute{H} + I - \acute{L}\,H(\acute{H}\,LH)^{-}\acute{H}\,] + ZLH(H/LH)^{-}\acute{H} - G\,(\acute{G}\,G)^{-}\acute{G}ZLH(\acute{H}\,LH)^{-}\acute{H}$$
$$\hat{Z}_{N.n} = G(\acute{G}\,G)^{-1}\acute{G}\,Z + GS(\acute{G}\,G)^{-1}H(\acute{H}\,(\acute{G}\,G)^{-1}H)^{-1}\acute{H} - G(\acute{G}\,G)^{-1}\acute{G}\,Z\,H(\acute{H}\,(\acute{G}\,G)^{-1}H)^{-1}\acute{H}\,(\acute{G}\,G)^{-1}$$
$$\hat{Z}_{N.n} = G(\acute{G}\,G)^{-1}\acute{G}\,Z + GS(\acute{G}\,G)^{-1}\acute{R}\,(R(\acute{G}\,G)^{-1}\acute{R})^{-1}R - G(\acute{G}\,G)^{-1}\acute{G}\,Z\,\acute{R}\,(R(\acute{G}\,G)^{-1}\acute{R}\,)^{-1}R\;(\acute{G}\,G)^{-1}$$
$$\hat{Z}_{N.n} = G\,(\acute{G}\,G)^{-1}\acute{G}\,Z + G(\acute{G}\,G)^{-1}\acute{R}\,(R(\acute{G}\,G)^{-1}\acute{R}\,)^{-1}R\acute{S} - G(\acute{G}\,G)^{-1}\acute{R}\,(R(\acute{G}\,G)^{-1}\acute{R}\,)^{-1}R\;(\acute{G}\,G)^{-1}\acute{G}\,Z$$

Where: $(\acute{G}\,G)^{-1}\acute{R}\,(R(\acute{G}\,G)^{-1}\acute{R}\,)^{-1}R = I_n$ (USING RIGHT AND LEFT INVERSE)

Where: $R^{-1}{}_{\text{left}} = (\acute{R}\,R)^{-1}\acute{R}$ and $\acute{R}^{-1}{}_{\text{right}} = R(\acute{R}\,R)^{-1}$

5.1. Two Sided Inverse

A two sided inverse of a matrix A is a matrix A^{-1} for which $AA^{-1} = I = A^{-1}A$. This is the inverse of A. When r = n = m; the matrix A has Full rank where n and m are the order of matrix A.

5.2. Left Inverse

Recall that A has full column rank if its columns are independent; i.e. if r = n. In this case the null space of A contains just the zero vector. The equation $Ax = b$ either has exactly one solution x or is not solvable.

The matrix $\acute{A}\,A$ is an invertible n by n symmetric matrix, so $(\acute{A}\,A)^{-1}\,\acute{A}\,A = I$, A^{-1} left $= (\acute{A}\,A)^{-1}\acute{A}$ is a left inverse of A [4].

Note that: AA^{-1} left is an m by m matrix which only equals the identity if $m = n$. A rectangular matrix can't have a two sided inverse because either that matrix or its transpose has a nonzero null space.

5.3. Right Inverse

If A has full row rank, then $r = m$. The null space of $\acute{A}$ contains only the zero vector; the rows of A are independent. The equation $Ax = b$ always has at least one solution; the null space of A has dimension $n - m$, so there will be $n - m$ free variables and (if n > m) infinitely many solutions. Matrices with full row rank have right inverses A^{-1} right with AA^{-1} right = I. The nicest one of these is $\acute{A}\;(A\,\acute{A}\,)^{-1}$. When times A to $\acute{A}\;(A\,\acute{A}\,)^{-1}$ is [4]

$$\because \hat{Z}_{N.n} = G\hat{\beta}_{OLS} + G(\acute{G}\,G)^{-1}\acute{R}\,(R(\acute{G}\,G)^{-1}\acute{R}\,)^{-1}r - G(\acute{G}\,G)^{-1}\acute{R}\,(R(\acute{G}\,G)^{-1}\acute{R}\,)^{-1}R\;\hat{\beta}_{OLS}$$
$$= G\hat{\beta}_{OLS} + G(\acute{G}\,G)^{-1}\acute{R}\,(R(\acute{G}\,G)^{-1}\acute{R}\,)^{-1}(r - R\;\hat{\beta}_{OLS})$$

$$\hat{Z}_{N.n} = X\hat{\beta}_{OLS} + X(\acute{X}X)^{-1}\acute{R}\,(R(\acute{X}X)^{-1}\acute{R}\,)^{-1}(r - R\ \hat{\beta}_{OLS}) = X\beta^{C}_{OLS} \quad (12)$$

The equation (12) indicate that:

$$\beta^{C}_{OLS} = \hat{\beta}_{OLS} + (\acute{X}X)^{-1}\acute{R}\,(R(\acute{X}X)^{-1}\acute{R}\,)^{-1}(r - R\ \hat{\beta}_{OLS}) \quad (13)$$

(This is the same result as (3))

6. Example of Unconstrained PCA and CPCA with Real Data

The data represent 1058 units of air condition that sold from July 2007 to March 2013 in an Egyptian company called Pure technology, we decomposed these units as The ISM frequency data on traditional vs. modern views is used [6], the data are as follows:

Table 1. *The count of sales units of air condition at different cases.*

No.	Sex	Cordon	Season	1.5 HP/b	2.25 HP/b	Hp/b	1.5 Hp/c	2.25 HP/c	3HP/c	TOTAL
1	M	Y	summer	17	6	13	52	32	26	146
2	M	Y	winter	3	0	0	3	1	2	9
3	M	Y	autumn	0	0	1	12	6	3	22
4	M	Y	spring	30	15	7	47	21	21	141
5	F	Y	summer	6	1	5	6	6	1	25
6	F	Y	autumn	0	0	0	0	3	1	4
7	F	Y	spring	1	0	1	4	0	3	9
8	C	Y	summer	0	0	0	0	2	6	8
9	C	Y	winter	0	0	0	2	0	1	3
10	C	Y	autumn	4	0	0	1	4	6	15
11	C	Y	spring	5	0	1	2	4	16	28
12	M	N	summer	20	15	11	29	26	29	130
13	M	N	winter	1	2	2	3	0	1	9
14	M	N	autumn	14	9	5	17	9	10	64
15	M	N	spring	45	13	11	37	29	21	156
16	F	N	summer	2	0	1	2	3	3	11
17	F	N	winter	0	0	1	4	3	1	9
18	F	N	autumn	1	1	1	5	1	3	12
19	F	N	spring	0	1	0	2	3	3	9
20	C	N	summer	2	1	2	1	8	28	42
21	C	N	winter	3	1	8	2	5	16	35
22	C	N	autumn	21	2	2	7	11	8	51
23	C	N	spring	9	5	4	12	28	62	120
				184	72	76	250	205	271	1058

(Collected from an Egyptian air condition Company called Pure Technology)

Where we make the cases constrained (G) is:

1. Sex of the client (M=Male, F=Female and C=company)
2. Cordon (place that the client live near from company or not) of the client (Y=Yes and N=No)
3. Season of the sale (summer, winter, autumn and spring).

And the variables constrained (H) is:

1. 1.5 HP/b represent the air condition with power 1.5 horse and it is hot and cold
2. 2.25 HP/b represent the air condition with power 2.25 horse and it is hot and cold
3. 3HP/b represent the air condition with power 3 horse and it is hot and cold
4. 1.5 HP/c represent the air condition with power 1.5 horse and it is cold
5. 2.25 HP/c represent the air condition with power 2.25 horse and it is cold
6. 3 HP/c represent the air condition with power 3 horse and it is cold

And the matrix G was as follows:

Table 2. *The cases constrained matrix* G.

M	F	C	Y	N	summer	winter	autumn	spring
G1	G2	G3	G4	G5	G6	G7	G8	G9
1	0	0	1	0	1	0	0	0
1	0	0	1	0	0	1	0	0
1	0	0	1	0	0	0	1	0
1	0	0	1	0	0	0	0	1
0	1	0	1	0	1	0	0	0
0	1	0	1	0	0	0	1	0

M	F	C	Y	N	summer	winter	autumn	spring
G1	G2	G3	G4	G5	G6	G7	G8	G9
0	1	0	1	0	0	0	0	1
0	0	1	1	0	1	0	0	0
0	0	1	1	0	0	1	0	0
0	0	1	1	0	0	0	1	0
0	0	1	1	0	0	0	0	1
1	0	0	0	1	1	0	0	0
1	0	0	0	1	0	1	0	0
1	0	0	0	1	0	0	1	0
1	0	0	0	1	0	0	0	1
0	1	0	0	1	1	0	0	0
0	1	0	0	1	0	1	0	0
0	1	0	0	1	0	0	1	0
0	1	0	0	1	0	0	0	1
0	0	1	0	1	1	0	0	0
0	0	1	0	1	0	1	0	0
0	0	1	0	1	0	0	1	0
0	0	1	0	1	0	0	0	1

(The data represent the constrained that found in cases, we get it from Table 1)

And the column constrained was constructed by combining between the power of the unit measuring by HP and the kind of this unit (cold only or cold and hot) and the matrix H was as follows:

Table 3. *The variables constrained matrix H.*

	1.5 HP	2.25 Hp	3 Hp	b	c
	H1	H2	H3	H4	H5
1.5 HP/b	1	0	0	1	0
2.25 HP/b	0	1	0	1	0
3Hp/b	0	0	1	1	0
1.5 Hp/c	1	0	0	0	1
2.25 HP/c	0	1	0	0	1
3HP/c	0	0	1	0	1

(The data represent the constrained that found in variables, we get it from Table 1)

We also use the profit of the unit as dependent variable to compare between OLS, PCA and CPCA the data is as follows:

Table 4. *The profit of the sales units of air condition at different cases.*

No.	Sex	cordon	season	1.5 HP/b	2.25 HP/b	3Hp/b	1.5 Hp/c	2.25 HP/c	3HP/c	TOTAL
1	M	Y	summer	6223	2474	5440	16947	11767	9918	52769
2	M	Y	winter	1050	0	0	335	210	849	2444
3	M	Y	autumn	0	0	440	4149	2055	1230	7874
4	M	Y	spring	11120	6040	2739	16161	7222	8461	51743
5	F	Y	summer	2124	449	2260	2000	1760	352	8945
6	F	Y	autumn	0	0	0	0	1150	410	1560
7	F	Y	spring	400	0	440	1399	0	1215	3454
8	C	Y	summer	0	0	0	0	188.31	2430	2618.31
9	C	Y	winter	0	0	0	-45	0	-123	-168
10	C	Y	autumn	-325	0	0	-75	4080	2176.68	5856.68
11	C	Y	spring	1265	0	440	350	1025	6560	9640
12	M	N	summer	7449	6094	3819	9323	9704	10608	46997
13	M	N	winter	450	898	849	1050	0	410	3657
14	M	N	autumn	5025	3252	2010	6025	2960	3461	22733
15	M	N	spring	9314	4390	4555	13214	10460	8516	50449
16	F	N	summer	1450	0	455	435	895	1230	4465
17	F	N	winter	0	0	440	1132	1199	449	3220
18	F	N	autumn	375	405	440	2150	210	1302	4882
19	F	N	spring	0	405	0	625	655	1315	3000
20	C	N	summer	1175	405	880	350	3105	10767	16682
21	C	N	winter	1050	455	3060	330	2195	6560	13650
22	C	N	autumn	2689	834	153.2	1987	2813	3228	11704.2
23	C	N	spring	889	1637	255	4116	5825	9510	22232
				51723	27738	28675	81958	69478.3	90834.7	350407.19

(Collected from an Egyptian air condition Company called Pure Technology)

The R programme version 2.4.1 is used to get the results, we found that there was found high correlation between variables, and the correlation matrix appear as follows:

Table 5. *Correlation matrix between the different types of the sold air condition.*

	1.5 HP/b	2.25 HP/b	3HP/b	1.5 Hp/c	2.25 HP/c	3HP/c
1.5 HP/b	1.00	0.85	0.73	0.80	0.78	0.41
2.25 HP/b		1.00	0.77	0.83	0.79	0.50
3 HP/b			1.00	0.84	0.84	0.52
1.5 Hp/c				1.00	0.85	0.44
2.25 HP/c					1.00	0.78
3HP/c						1.00

Unconstrained PCA done and the results shown that, the first two components explain 89.56% of the total information, i.e. when choosing two only component in case of reducing variables or removing multicolleinearity 10.44% of the total information will be lost.

Then the data matrix order become 23 x 2 and it consists from $Component\ 1\ and\ Component\ 2$ where:

$$\text{Component } 1 = -0.409\,z_1 - 0.423z_2 - 0.419z_3 - 0.426z_4 - 0.445z_5 - 0.313\,z_6.$$

$$\text{Component } 2 = 0.328\,z_1 + 0.221z_2 + 0.254\,z_4 - 0.243z_5 - 0.844z_6.$$

The scree plot indicate that the first component contribute more than 75% of the variation of the variables, where the second component approximately contribute with 10% as we shown in Figure 2 as follows:

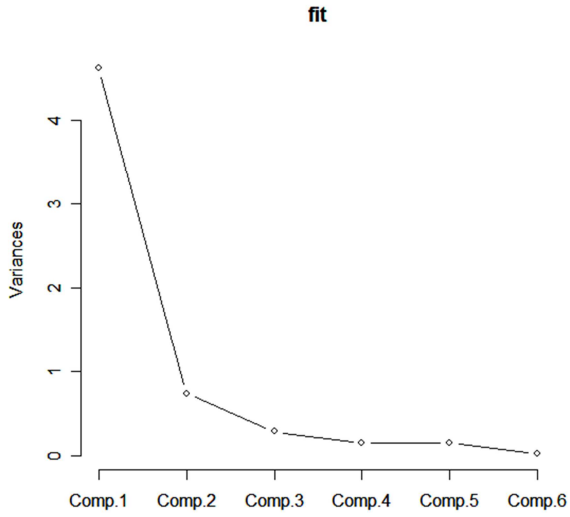

Figure 2. *Scree plot for the explained variance by each component in PCA method.*

CPCA shows that the first two components explain 98.45% of the total information, i.e. when choosing two only component in case of reducing variables or removing multicollinearity only 1.55% of the total information in the data will be lost. Then the data matrix order become 23 x 2 where and it consists from $Component\ 1\ and\ Component\ 2$ where:

$$\text{Component } 1 = -0.391z_1 - 0.418z_2 - 0.369z_3 - 0.429z_4 - 0.439\,z_5 - 0.399z_6.$$

$$Component\ 2 = 0.52\,z_1 + 0.295z_2 - 0.602\,z_3 + 0.237\,z_4 - 0.471z_6$$

The scree plot indicate that the first component contribute more than 80% of the variation of the variables, where the second component approximately contribute with 18% as we shown in Figure 3 as follows:

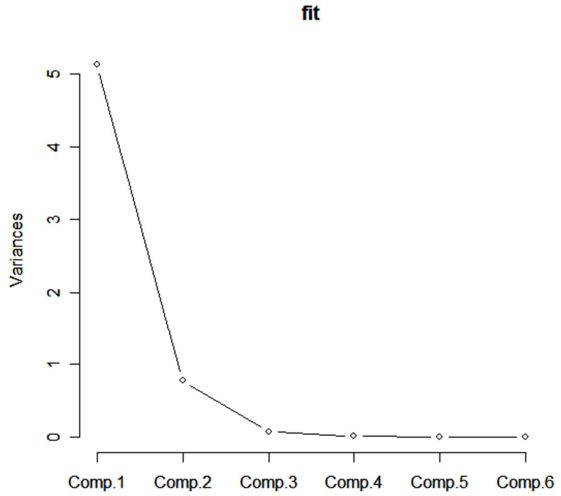

Figure 3. *Scree plot for the explained variance by each component in CPCA method.*

From the previous example, CPCA is better than PCA can be found, because the first have larger explaining of variation of the data where it used prior information $(G\ and\ H)$ about the data that contribute in explaining the total variation

7. Handling Data Differently

This section indicates the interpretation of the data as the ordinary least square, principal component analysis and the constrained principal component analysis when we applied multiple regression for the three cases and the estimation of the parameters for the OLS were as follows:

Table 6. *The estimation of the parameters for each air condition type.*

1.5 HP/b	2.25 HP/b	3Hp/b	1.5 Hp/c	2.25 HP/c	3HP/c
232.44	537.26	1,137.64	582.56	(221.59)	275.99

where 232.44 means that the profit will approximately arise to 230 pound when we sell one unit of 1.5HP/b, but -221.59 means that we will lose 220 pound when we sale one unit of 2.25HP/c i.e. we should to stop sell of this product. The confidence interval for these parameters were:

Table 7. *Confidence intervals for the estimation of the parameters for each air condition type.*

	2.50%	97.50%
1.5 HP/b	(7.54)	472.42
2.25 HP/b	(26.33)	1,100.86

	2.50%	97.50%
3Hp/b	525.74	1,749.54
1.5 Hp/c	350.64	814.48
2.25 HP/c	(756.89)	313.70
3HP/c	75.51	476.47

These intervals indicate that the profit of the 2.25HP/c product falls between -765 and 313 pound, and this with confidence level 95%, and the products 3HP/b, 1.5HP/c 3HP/c always achieve profit and did not make loss at any case. The values of the predicted value was as follows:

Table 8. *The predicted value of the regression.*

1	2	3	4	5	6	7	8
52,342.5	2,775.4	7,626.8	51,518.5	10,061.9	(388.8)	4,528.3	1,212.8
9	10	11	12	13	14	15	16
1,441.1	2,281.9	6,994.5	44,358.5	5,605.9	24,446.9	50,882.8	2,930.8
17	18	19	20	21	22	23	
3,079.1	5,426.5	1,865.6	9,815.0	14,808.8	12,079.4	27,226.5	

The sample number six means that the Females that live inside the cordon do not achieve profit in autumn season; this might need more advertisement for females in the cordon at autumn season. ANOVA table indicate that the all product are highly significant as follows:

Table 9. *ANOVA table for OLS.*

	Df	Sum sq	Mean sq	F value	Pr(>F)
1.5 HP/b	1	9.97E+09	9.97E+09	1.3849E+03	<0.0001***
2.25 HP/b	1	8.31E+08	8.31E+08	1.1549E+02	0.00001***
3 HP/b	1	8.64E+08	8.64E+08	1.2003E+02	0.00001***
1.5 Hp/c	1	2.59E+08	2.59E+08	3.5994E+01	0.0001**
2.25 HP/c	1	9.47E+07	9.47E+07	1.3160E+01	0.00208**
3HP/c	1	6.07E+07	6.07E+07	8.4358E+00	0.0099
Residuals	17	1.22E+08	7.20E+06		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

the estimation of the parameters for the PCA were as follows:

Table 10. *The estimation of the parameters for each linear combination of air condition type using PCA.*

z1.pc	z2.pc	z3.pc	z4.pc	z5.pc	z6.pc
(7,884.98)	1,420.00	(2,233.42)	54.57	(2,944.00)	(6,600.26)

The second combination z_2 is the better one because it achieves the most profit (1420 pound). The confidence interval for these parameters were:

Table 11. *Confidence intervals for the estimation of the parameters for each air condition type.*

	2.50%	97.50%
z1.pc	(11,548.38)	(4,221.58)
z2.pc	(7,707.42)	10,547.42
z3.pc	(16,846.42)	12,379.58
z4.pc	(20,008.45)	20,117.58
z5.pc	(23,146.88)	17,258.88
z6.pc	(52,965.28)	39,764.75

The interval also indicates that the first combination is the worst one, it always make loss. The second combination z_2 is the better one because it achieves less lost (7707 pound), but at the same time, it did not achieve highly profit as the last combination. The ANOVA table indicates that only the first combination is significant at the same time it did not make any profit and it was as follows:

Table 12. *ANOVA table for PC.*

	Df	Sum sq	Mean sq	F value	Pr(>F)
z1.pc	1	6.62E+09	6.62E+08	20.6215	0.0003***
z2.pc	1	3.46E+07	3.46E+07	0.1077	0.7467
z3.pc	1	3.34E+07	3.34E+07	0.1040	0.7510
z4.pc	1	1.06E+04	1.06E+04	0.0000	0.9955
z5.pc	1	3.03E+07	3.03E+07	0.0945	0.7622
z6.pc	1	2.90E+07	2.90E+07	0.0902	0.7676
Residuals	17	5.46E+09	3.21E+08		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

the estimation of the parameters for the CPCA were as follows:

Table 13. *The estimation of the parameters for each linear combination of air condition type using CPCA.*

z1.cpc	z2.cpc	z3.cpc	z4.cpc	z5.cpc	z6.cpc
-5.87E+03	4.48E+03	-2.43E+04	1.24E+05	4.13E+17	-1.72E+19

We also note that the second combination achieves profit 4483 while the fourth and the fifth combinations achieve more profits. The confidence interval for these parameters were:

Table 14. *Confidence intervals for the estimation of the parameters for each air condition type.*

	2.50%	97.50%
z1.cpc	-1.03E+04	-1.39E+03
z2.cpc	-1.04E+04	1.94E+04
z3.cpc	-3.39E+05	2.90E+05
z4.cpc	-4.23E+05	6.70E+05
z5.cpc	-6.91E+19	6.99E+19
z6.cpc	-6.87E+19	3.43E+19

The interval also indicate that the first combination is the worst one, it always make loss. The second combination z_2 is the better one because it achieve the less lost (1.044731e+04 pound), but at the same time it did not achieve highly profit as the fifth combination. ANOVA table indicate that only the first combination is significant at the same time it did not make any profit and it was as follows:

Table 15. *ANOVA table for CPC.*

	Df	Sum sq	Mean sq	F value	Pr(>F)
z1.cpc	1	4.49E+09	4.49E+09	10.39	.005***
z2.cpc	1	1.02E+08	1.02E+08	0.24	0.633
z3.cpc	1	3.35E+07	3.35E+07	0.08	0.784

	Df	Sum sq	Mean sq	F value	Pr(>F)
z4.cpc	1	2.46E+07	2.46E+07	0.06	0.814
z5.cpc	1	1.27E+06	1.27E+06	0.00	0.957
z6.cpc	1	2.15E+08	2.15E+08	0.50	0.490
Residuals	17	7.34E+09	4.32E+08		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

8. Numerical Example Using Bootstrap

To detect previous results of OLS, PC, and CPC using bootstrap with different sample size n, a numerical example has been made, The bootstrap method were applied to the original data at different sample size (20, 50, 100, 200, 500, 1000) with 1000 replications for each sample size, The results indicated the standard deviation sd, and the standard error se for the coefficients b of all types of air condition parameters at the three cases ordinary least square OLS, principal component PC, and constrained principal component CPC.

The results don't different in coefficients from the original data, where the OLS method indicate that all types of air conditions are made profit except 2.25Hp/c, but the PC and CPC methods indicate that all types don't made profit and these parameters don't have any effect with increasing the sample size, the results show also the standard error and the standard deviation of the variables decrease with the increasing of the sample size n, we note that sd and se for OLS < PC< CPC for all variables and this will make the confidence interval of CPC become more wide and that is mean that the decision making will be more accuracy, but the bootstrap method don't different from the original data in the proportion of the interpreting variance but it indicates that this proportion don't change with different of sample size.

9. Conclusion

From the results of the previous example we can conclude that Because of the high correlation between the variables, The OLS analysis refers only to the loss made by the fifth production z_5, while the PCA indicate that the significant first combination that contribute with 77 % in interpreting the total variation in the variables refers to a large loss falls between 11548.376 and 4221.576 where all six products made loss, that is the same information that indicated by the CPCA with 85% of interpreting the variation of the total information where it falls between 1.034412e+04 and 1.394876e+03. The previous results indicate that the company is not achieve any profit, it makes large loss, we should advice the owner to change his trade or deal with professional persons in the market to take advices from them and change his technique of management.

The previous example indicated that LM is an important statistical development in the last fifty years following GLM, PCA and CPCA in the last thirty years. This paper introduced a series of papers prepared within the framework of an international workshop. First, the LM and GLM has been discussed. Next, an overview of PCA has been presented as a tool for dealing with multicollinearity problem. then constrained principal component CPC has been shown. it was found that CPCA is better than PCA in solving the multicollinearity problem. then some of its special cases, related methods and example has been introduced to indicate the importance of CPCA and the different between PCA and CPCA. A real data of an air condition company has been used. The results show that CPC is more efficient than PC where it contribute more variation of the total variance that found in the data. then a bootstrap method has been done for the same data to indicate the behavior of the contributed variance with PCA and CPC with different sample sizes n. the results indicate that the sd and se of the combination variables decrease with increasing n and the proportion of the explained variance hasn't effected by n. finally ordinary least squares OLS estimator as a special case form CPCA has been shown.

References

[1] Batah, M. Özkale, M. and Gore, S. (2009) "Combining Unbiased Ridge and Principal Component Regression Estimators" Communications in Statistics - Theory and Methods, 38, 2201–2209.

[2] Guisan, A., Edwards, T. and Hastie, T. (2002) "Generalized linear and generalized additive models in studies of species distributions: setting the scene" Ecological Modelling, 157, 89-100.

[3] Gunst, R. F., Mason, R. L. (1977) "Biased estimation in regression: an evaluation using mean squared error" Journal of the American Statistical Association, 72, 616-682.

[4] *Hefferon, J. (2012) "Linear Algebra" Mathematics Department, Saint Michael's College.* http://joshua.smcvt.edu/linearalgebra.

[5] Hotelling, H. (1936) "Relations between two sets of variables" Biometrika, 28, 321–377.

[6] Hunter, M. and Takane, Y. (2002) "Constrained Principal Component Analysis: Various Applications" Journal of Educational and Behavioral Statistics, 27, 105- 145.

[7] Kruger, U., Zhang, J. and Xie, L. (2008) "Developments and Applications of Nonlinear Principal Component Analysis – a Review" Springer Berlin Heidelberg, 58, 1-43.

[8] Lambert, Z. V., Wildt, A. R. and Durand, R. M. (1988) "Redundancy analysis: An alternative to canonical correlation and multivariate multiple regression in exploring interset associations" Psychological Bulletin, 104, 282–289.

[9] Marquardt, D. (1970) "Generalized inverse, ridge regression, biased linear estimation and nonlinear estimation" Technometrics, 12, 591-612.

[10] Massy, W. F. (1965) "Principal component regression in explanatory statistical research" Journal of the American Statistical Association, 60, 234-256.

[11] MIT (2011) "Left and right inverses; pseudoinverse" Massachusetts Institute of Technology (MIT), OpenCourseWare Linear Algebra, 1-4.

[12] Pearson, K. (1901) "On lines and planes of closest fit to systems of points in space" Philosophical Magazine, 2, 559–572.

[13] Takane, Y. (1997) "CPCA: A Comprehensive Theory" Department of Psychology, McGill University Montreal, Quebec H3A lB1, CANADA, 35- 40.

[14] *Takane, Y. (2014) "Constrained Principal Component Analysis and Related Techniques" CRC Press Taylor and Francis Group.*

[15] Takane, Y. and Hunter, M. (2001) "Constrained Principal Component Analysis: A Comprehensive Theory" Applicable Algebra in Engineering, communication and computing, 12, 391–419.

[16] Takane, Y., Kiers, H. A. L., and de Leeuw, J. (1995) "Component analysis with different sets of constraints on different dimensions" Psychometrika, 60, 259-280.

[17] Takane, Y., and Shibayama, T. (1991) "Principal Component Analysis With External Information on Both cases and Variables" Psychometrica Mcgill University, 56, 97-120.