On Constrained Principal Component Analysis (CPCA) with Application on Bootstrap

Salah M. Mohamed* Alaa A. Abd Elmegaly **

Abstract

Linear model (LM) provide the advance in regression analysis, where it was considered an important statistical development of the last 45 years, following general linear model (GLM), principal component analysis (PCA) and constrained principal component analysis (CPCA) in the last 25 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, 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.

Key words: 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),

$$Y_{n.1} = X_{n.p}\beta_{p.1} + \epsilon_{n.1} \quad (1)$$

*Assistant Professor, Department of Applied Statistics and Econometrics, Institute of Statistical Studies and Research, Cairo University.

**Ph.D. student, Department of Applied Statistics and Econometrics, Institute of Statistical Studies and Research, Cairo University.

components. Mathematically, PCA depends on the eigen decomposition of positive semi definite matrices and on the singular value decomposition SVD of rectangular matrices (**Kruger** and et al, 2008) and. In case of multicollinearity problem, the researchers used another forms to estimate the parameters like principal component regression PCR (**Batah** et al, 2009), 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. (**Massy**, 1965), (**Marquardt**, 1970), and (**Gunst** and Mason, 1977) used the orthogonal matrix T in the GLM to get the PCR estimator for β as:

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

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

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

Where $\Lambda_r = \acute{T}_r \acute{X} X T_r$ 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 Λ_{p-r} matrix are the remaining $p - r$ eigenvalues.

The PCR estimator for β can be written as

$$\widehat{\beta}_{PC} = T_r\ (\acute{T}_r \acute{X}\, X T_r)^{-1}\ \acute{T}_r \acute{X}\, Y \tag{6}$$

Expectation and variance:

$$\mathbb{E}(\widehat{\beta}_{PC}) = T_r \acute{T}_r \beta = I_r \beta\ (biased)$$

$$var(\widehat{\beta}_{PC}) = \sigma^2\ T_r\ (\Lambda_r)^{-1} \acute{T}_r$$

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 (**Takane** and **Shibayama**, 1991).

The constrained principal component model is:

$$Z_{N.n} = G_{N.p} M_{p.q} \acute{H}_{q.n} + B_{N.q} \acute{H}_{q.n} + G_{N.p} C_{p.n} + E_{N.n} \tag{7}$$

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 $\sigma^2 I$. (**Takane**, 2014) estimated the unknown matrices of parameter as:

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

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

$$\widehat{C} = (\acute{G}\,KG)^{-}\acute{G}\,KZ\acute{Q}_{H/L}LL^{-} \tag{10}$$

4. Some Special cases and related methods of CPCA:

(**Takane**, 2014) introduced about 20 special cases and related techniques for CPCA as PCA, CANO, and Redundancy analysis (RA) the next part indicates some of them and illustrate the assumptions that lead each case to CPCA.

4.1 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 $L = I_n$ (**Takane** and **Hunter**, 2001).

4.2 Canonical correlation analysis CANO, proposed by (Hotelling, 1936), 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 $Z = (\acute{G}\,G)^{-}\acute{G}\,H(\acute{H}\,H)^{-}$, $K = \acute{G}\,G$, $L = \acute{H}\,H$, $G = I$, and $H = I$.

4.3 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 (**Lambert** et al. 1988). RA follows from CPCA by setting $H = I, K = I\ and\ L = I$.

4.4 Correspondence Analysis (CA) When both G and H consist of dummy coded categorical variables, CANO specializes in correspondence analysis CA of a probability table $F = \acute{G}\,H$.

4.5 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 (**Takane**, 2014).

4.6 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.

4.7 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).**

- **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.

- **Left inverse**

Recall that A has full column rank if its columns are independent; i.e. if r = n. In this case the nullspace 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 (**Hefferon**, 2012).

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.

- **Right inverse**

If A has full row rank, then **r = m**. The nullspace 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 nullspace 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 I (**Hefferon**, 2012).

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

$$= G\hat{\beta}_{OLS} + G(\acute{G}G)^{-1}\acute{R}(R(\acute{G}G)^{-1}\acute{R})^{-1}(r - R\,\hat{\beta}_{OLS})$$

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

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, that found in (**Hunter** and **Takane**, 2002), the data are 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
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:

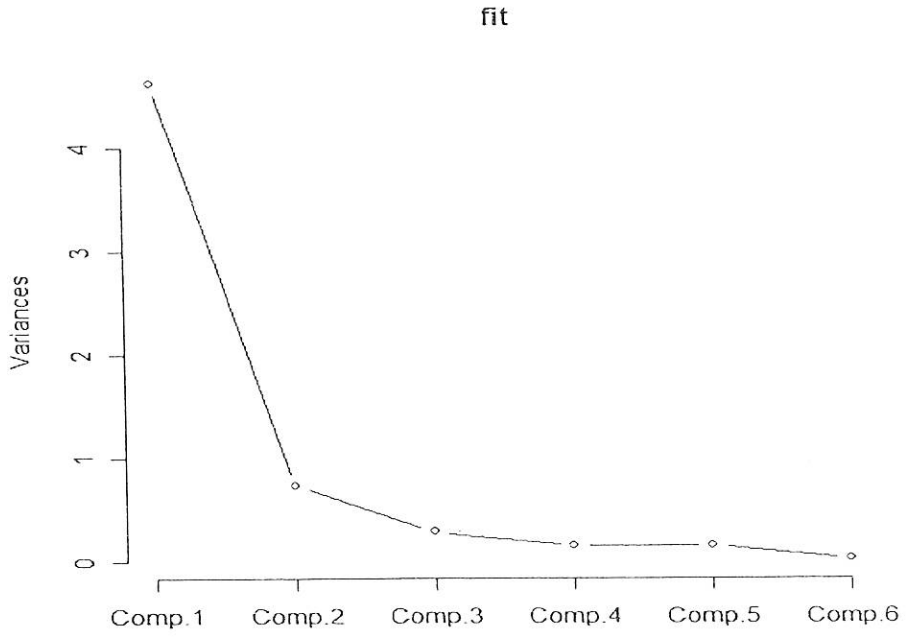
Then the data matrix order become 23 x 2 and it consists from *Component* 1 *and Component* 2 where:

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

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 (1) as follows:

Figure (1): 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:

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 (2) as follows:

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
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***
3Hp/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 ' '					

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:

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	0.005***
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
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 ' '					

Conclusion:

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.

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

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 1B1, 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.