Advances in Information Mining
ISSN: 0975–3265 & E-ISSN: 0975–9093, Vol. 3, Issue 1, 2011, pp-26-32
Available online at http://www.bioinfo.in/contents.php?id=32

ANALYSIS OF FACTOR BASED DATA MINING TECHNIQUES

ABHISHEK TANEJA[1]* AND CHAUHAN R.K.[2]
[1]Department of Computer Applications, DIMT, Kurukshetra, India
[2]Department of Computer Sc. & Applications, Kurukshetra University, India
*Corresponding Author: Email- taneja246@yahoo.com

Received: September 01, 2011; Accepted: September 19, 2011

Abstract- Factor analysis, which is a regression based data mining technique, used to represent a set of observed variables in terms of common factors. This paper explores the key properties of three factor based techniques viz. principal component regression, generalized least square regression, and maximum likelihood method and study their predictive performance on theoretical as well as on experimental basis. The issues such as variance of estimators, normality of distributed variance of residuals, effect of multicollinearity, error of specification, and error of measurement are addressed while comparing their predictive ability.
Keywords - Factor Analysis, Principal Component Analysis (PCA), Generalized Least Square Regression (GLS), Maximum Likelihood Regression (MLR), Data Mining
.

Introduction

Factor analysis is set of techniques used to find out the underlying constructs which influence the responses on a number of measured variables. All the techniques are based on common factor model, represented in figure 1.

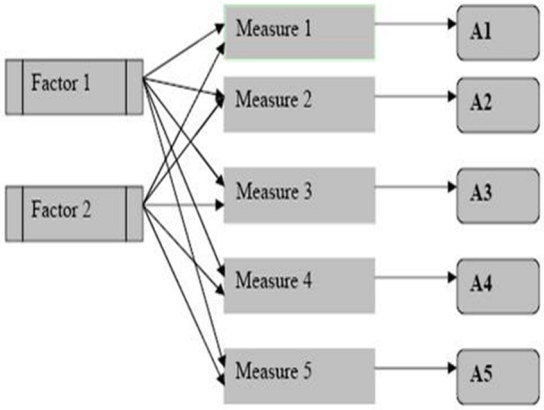


Fig. 1-The Factor Model

The factor model illustrates that each observed prediction (from measure 1 to measure 5) is influenced by the underlying latent variables/common factors (factor 1 and factor 2) and to some extent by underlying unique factors (A1 to A5). The common factors are latent variables which explain why a number of variables are correlated with each other- it is because they have one or more factors in common [1].

Factor analysis is essentially a one-sample method [2]. For example, we presume a sample $\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{X}_n$ from a homogeneous population with mean vector μ and covariance matrix $\sum$. The factor analysis model represents each variable as a linear combination of underlying *common factors* f_1, f_2, . . . , f_m, with an associated residual term to account for that part of the variable that is unique. For X_1, X_2, X_p in any observation vector **X**, the model is as follows:

$$X_1 - \mu_1 = \lambda_{11} f_1 + \lambda_{12} f_2 + \cdots + \lambda_{1m} f_m + \varepsilon_1$$
$$X_2 - \mu_2 = \lambda_{21} f_1 + \lambda_{22} f_2 + \cdots + \lambda_{2m} f_m + \varepsilon_2$$
$$\ldots$$
$$X_p - \mu_p = \lambda_{p1} f_1 + \lambda_{p2} f_2 + \cdots + \lambda_{pm} f_m + \varepsilon_p.$$

Preferably, m should be significantly smaller than p; or else we have not achieved a prudent explanation of the variables as functions of a few underlying factors [3]. The *f's* in equations above as random variables that produce the *X*'s. The coefficients λ_{ij} are called *loadings* and serve as weights. They show how every X_i independently depends on the *f* 's. With suitable assumptions, λ_{ij} indicates the significance of the *j*th factor f_j to the *i*th variable X_i and can be used in interpretation of f_j. We describe or interpret f_2, for example, by examining its coefficients, λ_{12}, λ_{22}, λ_{p2}. The larger loadings relate f_2 to the corresponding *X*'s. From these *X*'s, we infer a meaning or description of f_2. After estimating the λ_{ij} 's, it is hoped they will partition the variables into groups corresponding to factors. At the very first it appears that the multiple linear regression and factor analysis are similar techniques but there are essential differences. For example, firstly f's in above equations are unobserved, secondly equations above represents one observational vector, whereas multiple linear regression depicts all n observations.

Principal Component Analysis

Principal components analysis (PCA) seeks to explain the correlation structure of a set of predictor variables using a smaller set of linear combinations of these variables. These linear combinations are called *components*. The total variability of a data set produced by the complete set of *m* variables can often be accounted for primarily by a

smaller set of *k* linear combinations of these variables, which would mean that there is almost as much information in the *k* components as there is in the original *m* variables. PCA is heavily used for dimensionality reduction also where the analyst can replace the original *m* variables with the *k<m* components, so that the operational data set now consists of *n* records on *k* components rather than *n* records on *m* variables.

Generalized Least Square

Assume, $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$, a linear regression model, where **y** is a *n* X 1 vector of observations on a dependent variable, **X** is a *n* X *k* matrix of independent variables of full column rank, β is a *k* X 1 vector of parameters to be estimated, and **u** is a X 1 vector of residuals [4]. Here V satisfy the Gauss-Markov Theorem, if

A1 $E(\mathbf{u}|\mathbf{X}) = \mathbf{0}$ (i.e., the residuals have conditional mean zero), and

A2 $E(\mathbf{uu'}|\mathbf{X}) = \sigma^2 \Omega$, where $\Omega = \mathbf{I}_n$, is a *n* X *n* identity matrix (i.e., conditional on the **X**, the residuals are independent and identically distributed or ''iid'' with conditional variance σ^2), then the ordinary least squares (OLS) estimator $\hat{\beta}_{OLS} = (\mathbf{X'X})^{-1}\mathbf{X'y}$ with variance-covariance matrix $V(\hat{\beta}_{OLS}) = \sigma^2 (\mathbf{X'X})^{-1}$ (1) satisfies best linear unbiased estimator (BLUE) properties of β; (2) a consistent estimator of β (i.e., as $n \to \infty$ 1, $Pr[|\hat{\beta}_{OLS} - \beta| < \varepsilon] = 1$, for any $\varepsilon > 0$, or plim $\hat{\beta}_{OLS} = \beta$).

If A2 fails to hold (i.e., $\Omega \neq \mathbf{I}_n$, where Ω is a positive definite matrix but not equal to $\mathbf{I}_n$), then $\hat{\beta}_{OLS}$ remains unbiased, but no longer ''best'', and remains consistent. Relying on $\hat{\beta}_{OLS}$ when A2 doesn't hold risks faulty inferences; without A2, $\hat{\sigma}^2(\mathbf{X'X})^{-1}$ is a biased and inconsistent estimator of $V(\hat{\beta}_{OLS})$, meaning that the estimated standard errors for $\hat{\beta}_{OLS}$ are wrong, invalidating inferences and the results of hypothesis tests. Assumption A2 often fails to hold in practice: e.g., (1) when pooling across disparate units generates disturbances with different conditional variances (*heteroskedasticity*); (2) an analysis of time series data generates disturbances that are not conditionally independent (*serially correlated disturbances*).

When A2 does not hold, it may be possible to implement a *generalized least squares* (GLS) estimator that is BLUE (at least asymptotically). For instance, if the researcher knows the exact form of the departure from A2 (i.e., the researcher knows Ω) then the GLS estimator $\hat{\beta}_{GLS} = (\mathbf{X'}\Omega^{-1}\mathbf{X})^{-1}\mathbf{X'}\Omega^{-1}\mathbf{y}$ is BLUE, with variance-covariance matrix $\sigma^2 (\mathbf{X'}\Omega^{-1}\mathbf{X})^{-1}$. Note that when A2 holds, $\Omega = \mathbf{I}_n$ and $\hat{\beta}_{GLS} = \hat{\beta}_{OLS}$ (i.e., OLS is a special case of the more general estimator)[5,6].

Maximum Likelihood Regression

Maximum likelihood regression (MLR) is a method that finds the most likely value for the parameter based on the data set collected [7]. MLR is by far is the most popular method of parameter estimation and a vital tool for many statistical modeling techniques particularly in non-linear modeling and non-normal data.

The theory of MLR states that the desired probability distribution be the one that makes the observed data most likely and that can be obtained by finding the value of the parameter vector that maximizes the likelihood function F(w). The resulting parameter which is found by searching the multidimensional parameter space is called as MLR estimate, denoted by $F_{MLR}=(F_{1,MLR},\ldots\ldots\ldots F_{K,MLR})$.

For computational convenience it is prescribed that MLR estimate is obtained by maximizing the log likelihood function ln F(w). Assuming the log likelihood function, ln F(w), is differentiable, if W_{MLR} exists, and must satisfy the following likelihood equation.

$$\frac{\partial \ln F(w)}{\partial w_i} = 0$$

At $w_i = w_{i,MLR}$ for all i=1,….,k. The likelihood equation represents the necessary condition for the existence of an MLR estimate. The other condition that need to be satisfied to ensure that $F(W_{MLR})$is maximum and not a minimum. Formally the above discussion can be described as:

If $y_1, y_2, \ldots\ldots\ldots, y_n$ is a random sample of size n from a discrete or continuous probability density function, f_Y (yi; θ), where θ is an unknown parameter then the likelihood function is written

$$L(\theta) = \prod_{i=1}^{n} f_Y(y_i;\theta)$$

Methodology

To intra compare the three factor based techniques for their predictive performance we have chosen three unique datasets. The three datasets viz., marketing, bank, and Parkinson tele-monitoring datasets have been procured from [8], [9], and [10, 12] respectively. All the datasets included in this study are unique and are having many missing values. Missing values should be filled prior to using it for modeling. Missing values should be filled in such a manner to avoid biasness and keeping the patterns available in the dataset intact. They can be best

filled using regression based techniques but this requires many resources. So, in this paper we used filling missing values by the mean of that column. This is not a serious problem as the objective of this study is to compare the performance of factor based techniques not to evaluate the results after the deployment of the model. All the datasets have been preprocessed by taking natural log of all the instances or by taking normalization by z-score normalization just to make them linear. After standardizing three datasets, they are divided into two parts, taking 70% observations as the "training set" and the remaining 30% observations as the "test validation set" [11]. For each data set training set is used to build the model and various methods of that technique are employed. The models build are then evaluated on the basis of ten model fitness criteria.

Interpretation

Table I, contains the experimental results of three factor based techniques. With reference to this table in marketing dataset, the value of R^2 and Adj.R^2, of maximum likelihood model was found with good explanatory power i.e., 0.589, which is higher than both PCR and GLS model.

On the behalf of this explanatory power value we can say that among all methods of factor analysis, maximum likelihood model was found best method for data mining purpose, since almost 59% change in variation in dependent variable was explained by independent variables. But 0.589 value of explanatory power is although good yet requires another regression model than factor analysis model for reporting data set, since 0.41 means 41% of the total variation was found unexplained. So, within factor analysis techniques maximum likelihood model was found best but not up-to the mark. Value of R^2 suggest for using another regression model. R^2 can also be estimated through the following notations:

$$R^2 = \frac{ESS}{TSS}$$

TSS = Explained Sum Square(ESS)+ Residual Sum Square(RSS)

The Adj. R^2 is maximum again in maximum likelihood i.e. 0.576, adjusted for inclusion of new explanatory variable less than R^2. The 58% variation was captured due to regression, it explains the overall goodness of fit of the regression line to marketing dataset due to use of factor analysis.

So, on the behalf of first order statistical test (R^2), we can conclude that maximum likelihood model of factor analysis technique is better than multiple regression technique due to explanatory power.

Mean Square Error (MSE) criteria is a combination of unbiased-ness and the minimum variance property. An estimator is a minimum MSE estimator if it has smallest MSE, defined as the expected value of the squared differences of the estimator around the true population parameter b. MSE($\hat{b}$) =E($\hat{b}$ -b)^2 . It can be proved that it is equal to

$$\text{MSE}(\hat{b})\text{'s} = \text{Var}(\hat{b})\text{'s} + \text{bias}^2(\hat{b})$$

The MSE criteria for unbiased-ness and minimum variance were found maximum in case of maximum likelihood model of factor analysis. It signifies that Maximum likelihood MSE is more than all other model's MSE, which further means that under this model of factor analysis of marketing dataset there is more unbiased-ness and more variance.

The more variance also increases the probability of biased-ness and gives unexpected explanatory power like R^2 in marketing dataset.

The intra comparison of three techniques of factor analysis generated that in factor analysis models MSE is significantly different and contradictory which signifies that under factor analysis all b's are biased but with large variance. Due to large variance in factor analysis techniques the probability value of unbiased-ness increases that generates a contradictory result about the explanatory power of the factor analysis methods. But factor analysis methods may have questionable values of MSE, due to this reason new measure of MSE that is RMSE (root mean square error) was used in the study.

RMSE was found considerably similar in all the three techniques. Due to less variation in RMSE of three models of factor analysis of marketing dataset it can be stated that three techniques have equal weights for consideration.

A common measure used to compare the prediction performance of different models is Mean Absolute Error (MAE).

If Y^p be the predicted dependent variable and Y be the actual dependent variable then the MAE can be computed by

$$\text{MAE} = \frac{1}{n}\frac{\sum\left|Y - Y^p\right|}{Y}$$

In marketing dataset MAE was found less under PCR model, which is less than GLS and maximum likelihood model. MAE signifies that PCR model under factor analysis techniques give better prediction than other model.

Under factor analysis marketing dataset MAE in all models was found considerably similar but higher than required, therefore we can say factor analysis models for such kind of datasets generate poor prediction performance.

The diagnosis index of multi collinearity was found significantly below 100 under factor analysis methods in marketing dataset, which means there is no scope for high and severe multi collinearity. In case maximum likelihood of same dataset condition number was found lowest than PCR and GLS technique. This means maximum likelihood is better technique to diagnosis the effect of multi collinearity. But in marketing dataset all techniques were found with less multi collinearity in regressors than severe level of multi collinearity.

The F value in case of marketing dataset was found more in case of GLS than rest of techniques, which signifies that overall regression model is significantly estimated but GLS model of factor analysis technique was found high F

corresponding to its dF which means overall significance of the regression model was up-to the mark in case of GLS method.

A Scree plot, which is a simple line segment plot that shows the fraction of total variance in the data as explained or represented by principal components. The principal components are ordered, and by definition are therefore assigned a number label by decreasing order of contribution to total variance. The principal component with the largest fraction contribution is labeled with the label name from the preference file. Such a plot when read left to right across the abscissa can often show a clear separation in the fraction of total variation where the most important components cease and least.

In marketing dataset of the study scree plot shows that number of principal componets should be six since the critical eigen value of components is one here and beyond six components eigen value would be less than one which signifies that components after this eigen value have less fraction in the total variation (see figure 2).

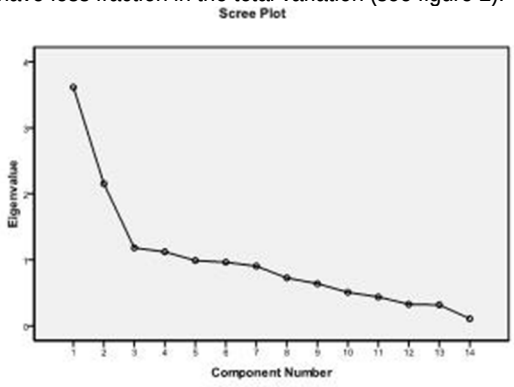


Fig. 2-Scree plot of PCR on marketing dataset

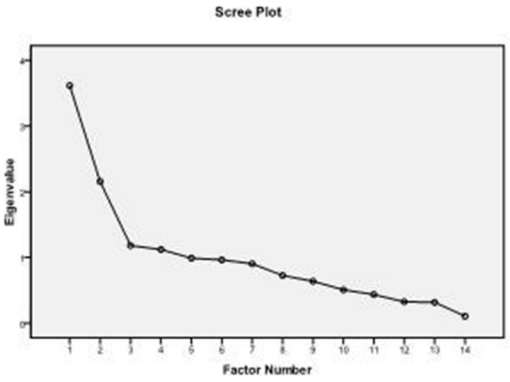


Fig. 3-Scree plot of Maximum Liklihood on marketing dataset

Under maximum likelihood model of factor analysis on marketing dataset, scree plot was found to suggest five number of components for getting fraction in total variation (figure 3).

In GLS method of factor analysis for marketing dataset scree plot prescribed number of components are five which is similar to maximum likelihood method of factor analysis (figure 4).

All the three techniques of factor analysis models on bank dataset generated higher value of both R^2 and adjusted R^2, which signifies that the explanatory power of factor analysis in case of bank dataset is more as compared to marketing dataset.

The MSE criteria for unbiasness and minimum variance for all parameters is found increasing under factor analysis, but all models of factor analysis are found with low unbiasness and variance. It means all the technique parameters are significant.

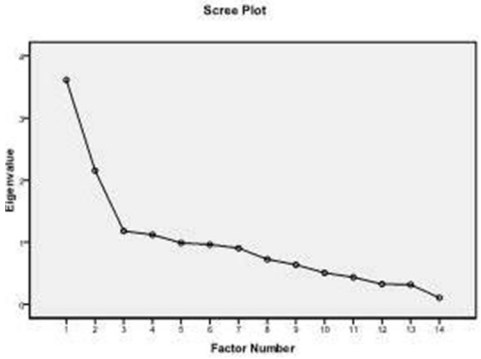


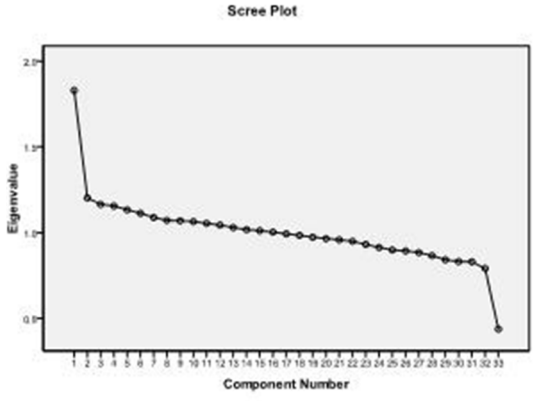
Fig. 4-Scree plot of GLS on marketing dataset

The RMSE is also very similar to marketing dataset. It is satisfactory and up-to the mark in all the three techniques.

The prediction power of the regression model is also found good fit in all factor analysis models. Modified coefficient of efficiency was found low in case of factor analysis model in case of bank dataset, since this dataset does not satisfy the center limit theorem due to constant number of variables.

In case bank dataset the diagnosis index of multi-collinearity was found almost similar in all the three techniques of factor analysis, which signifies that three techniques are equally powerful to identify multi-collinearity problem.

For the bank dataset scree plot of PCR suggested that number of principal components should be thirteen for thirty three variables (figure 5).



Fig, 5-Scree plot of PCR on bank dataset

Under GLS and maximum likelihood method again scree plot was found to suggest that number of components should be thirteen for thirty three variables (figure 6 to 10). On the basis of this we can say that all method of factor analysis are similar capability to extract variation out of total variation. The impact of residual/random error has

been minimized which further supports the BLUE properties of regression modeling.

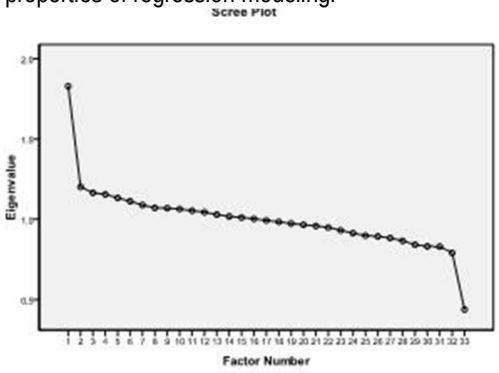


Fig. 6-Scree plot of Maximum Liklihood on bank dataset

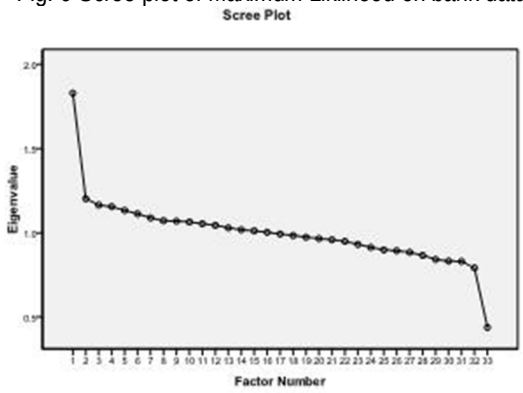


Fig. 7-Scree plot of GLS on bank dataset

In GLS model of factor analysis R^2 and Adj. R^2 was found to have around 67% and 56% respectively, which is considerably sufficient for satisfactory explanatory power of the model. This is due to no intrapolation.

The MSE value is low in all the three models of factor analysis but it is lowest in case of GLS which signifies that GLS technique is better technique for the extraction of structural parameters with unbiasness and low variance.

RMSE value also shows similar pattern in all the three models as MSE, which signifies same consideration for unbiasness and variance.

The prediction power (MAE) of GLS model of factor analysis was found maximum and considerably higher than PCR and maximum likelihood.

The modified coefficient of efficiency for getting efficiency in the model was found maximum in case of GLS due to successful implementation of center limit theorem.

The multi-collinearity extraction index was found more or less similar in all the three models of factor analysis, which signifies that all are similar as far as diagnosing multi-collinearity is concerned.

The significance of overall model was found highest in case of GLS, which signifies that overall regression model is better estimated in this model.

Parkinson dataset's PCR and maximum likelihood method, scree plot were found to suggest that number of principal components should be four but GLS method was found to suggest five components for regression modeling.

For the appropriate regression modeling under factor analysis scree plot plays an important role by supporting the assumption of BLUE. The fraction of variation explained out of total variation can be judged through number of components and which further can be decided by scree plot.

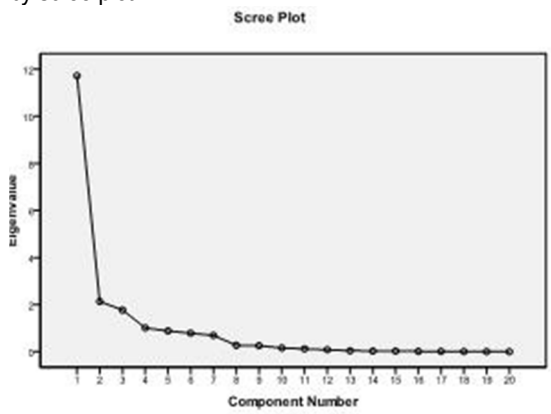


Fig. 8-Scree plot of PCR on Parkinson dataset

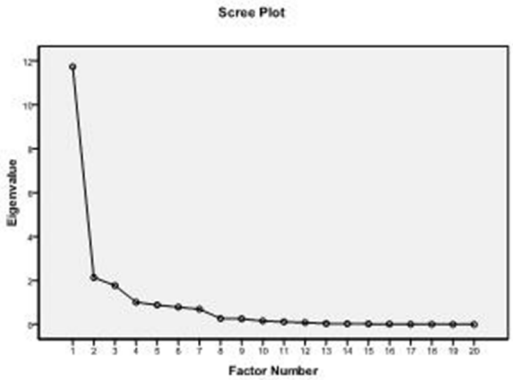


Fig.9-Scree plot of Maximum Liklihood on Parkinson dataset

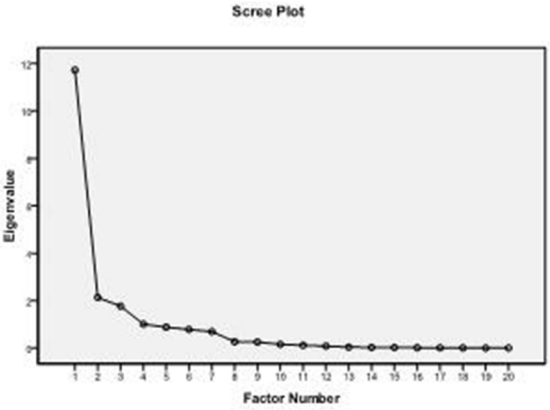


Fig. 10-Scree plot of GLS on Parkinson dataset

So, on the basis of above explanation we can say that scree plot of eigen values and component number is a bench mark contribution for up-to the mark regression modeling(regression modeling with desirable properties of its coefficients and satisfaction of usual assumption).

Conclusion

There are various linear techniques to extract the structural parameters of the regression model and factor analysis is also one of them which comprises of three major techniques i.e. principal component regression,

GLS (Generalized Least Square) and maximum likelihood method.

These methods or techniques yield good or desirable estimates of parameters if and only if when they are fitted to the unique datasets (which are randomly selected). These techniques entail the linearity in parameters and linearity in variables. In our s tudy of three datasets there is linearity in parameters of regression model fitted for randomly selected data sets not in the variables but at least one condition of linearity has been satisfied in this regard.

With the linearity of parameters of regression model fitted under factor analysis techniques for study, the assumption of least variance of regression model should be satisfied. Out of the three techniques of factor analysis PCR technique is considered as best for least variance and with low effect of multi-collinearity. For large datasets PCR has been found to support results in accordance to theoretical ground in contrast to maximum likelihood which is found to support small datasets like marketing.

Theoretically GLS and maximum likelihood techniques of factor analysis are considered to have unbiased ness but with large variances. In our study of three datasets these two also have performed well to satisfy Gauss Markov Theorem (least variance property) than PCR.

In all linear techniques factor analysis techniques performed well with least variance of residual and least variance of estimators but this performance differs from one dataset to another dataset.

The ranking of the techniques on the basis of theoretical and studied datasets can be generalized in the table II.

References

[1] Kim Jae-on., Mueller Charles W. (1978) "Introduction to Factor Analysis-What it is and how to do it.", Sage Publications, Inc.

[2] Rencher C. Alvin (2002) "*Methods of Multivariate Analysis" 2nd Edition, Wiley Interscience*.

[3] Kim Jae-on and Charles W. Mueller (1978). *Factor analysis : statistical methods and practical issues. Beverly Hills, Calif.: Sage Publications.*

[4] Chipman J.S. (1979) *Econometrica* 47:115--128.

[5] Amemiya Takeshi (1985) *Advanced Econometrics. Cambridge: Harvard University Press.*

[6] Cochrane D. and Orcutt G.H. (1949) *Journal of the American Statistical Association,* 44:32--61.

[7] Kleinbaum David G., Lawrence L. Kupper and Keith E. Muller (1998) *Applied Regression Analysis and Other Multivariate Methods. Belmont, California: Duxbury Press*.

[8] http://www-stat.stanford.edu/~tibs/ElemStatLearn/,

[9] http://www.cs.toronto.edu/~delve/data/bank/desc.htm

[10] http://archive.ics.uci.edu/ml/datasets.html

[11] Myatt J. Glenn (2007) *Making Sense of Data-A practical guide to exploratory data analysis and data mining, New Jersy: Wiley-Interscience*.

[12] Little M.A., McSharry P.E., Roberts S.J., Costello D.A.E., Moroz I.M. (2007) *BioMedical Engineering OnLine*, 6:23.

Table I

	Methods	MSE	MAE	CN	No. of variables	R Square	Adj. R Square	RMSE	F-Value (df, No. of Observations)	Modified Coefficient of efficiency	Test of normality
FACTOR ANALYSIS (*MARKETING DATASET*)	PCR	0.756	3.67	12	13 *(with four components)*	0.584	0.56	0.8694	323.65 (13,4819)	5.754	0.6654
	MAXIMUM LIKLIHOOD	0.775	3.98	9.78e+9	13	0.589	0.576	0.8803	367.455 (13,4819)	5.9876	0.6792
	GLS	0.746	3.998	11	13	0.587	0.573	0.8602	386.78 (13,4819)	5.7685	0.6776
FACTOR ANALYSIS (*PARKINSON DATASET*)	PCR	0.456	0.67	7.87e+7	19 *(with six components)*	0.63	0.51	0.6749	543.5 (19,4112)	8.56	0.87
	MAXIMUM LIKLIHOOD	0.582	0.655	7.10e+7	19	0.64	0.54	0.763	513.65 (19,4112)	9.38	1.73
	GLS	0.398	1.677	6.54e+6	19	0.67	0.56	0.63	665.45(11, 4112)	11.09	1.96
FACTOR ANALYSIS (*BANK DATASET*)	PCR	0.643	0.58	8.86e+8	33 *(with six components)*	0.74	0.69	0.80	654.45 (34,3150)	0.0544	0.6758
	MAXIMUM LIKLIHOOD	0.665	0.598	8.75e+8	33	0.728	0.684	0.815	675.65 (34,3150)	0.0546	0.0754
	GLS	0.678	0.612	8.74e+8	33	0.715	0.682	0.823	688.45 (34,3150)	0.0568	0.0543

Table II

On The Basis Of→		Theoretically	Studied Datasets
1.	Least Variance of Estimators	PCR and GLS	PCR and GLS
2.	Normality of Distributed Variance of Residual	PCR and Maximum Likelihood	GLS and PCR
3.	Least Effect of MultiCollinearity	PCR and GLS	PCR and Maximum Likelihood
4.	Error of Specification	GLS and Maximum Likelihood	GLS and Maximum Likelihood
5.	Error of Measurement	GLS and Maximum Likelihood	PCR and GLS