

## **A FACTOR ANALYSIS APPROACH IN PERFORMANCE ANALYSIS OF T-20 CRICKET**

**Sujeet Kumar Sharma**

Department of Operations Management and Business Statistics, CEPS,  
Sultan Qaboos University, Oman  
Email: drsujeet@squ.edu.om

(Received October 03, 2012)

### **Abstract**

This paper investigates the systematic covariation among various dimensions pertaining to batting and bowling capabilities of T20 cricket using the advanced statistical technique of factor analysis. The real dataset of 85 batsmen and 85 bowlers has been considered from Indian Premier League (IPL) 2012 for analysis purpose. The findings of this study show that batting capability dominates over bowling capability. This conclusion coincides with the general opinion of several cricketing enthusiasts and experts; however till date, there is no research study, to the best of the author's knowledge that provides statistical evidence to support this notion.

**Key words:** Factor analysis, batting, bowling, T20, and cricket

### **1. Introduction**

Cricket, a game popular in the majority of the British Commonwealth Countries as well as some other countries, is played in a standard format called a test match for a long period. The test match is a two innings per team contest that is played over five days. The monotony of the game owing to the long duration and inconclusiveness of end result referred to as a *draw* in cricketing jargon, newer formats evolved. The newer format shortened the duration to one where each team plays one innings with limited number of overs. This format was commercially successful and spectators enjoyed shorter version of the cricket. To those who are not familiar with the game of cricket, Preston and Thomas (2000) summarize the important concepts of limited overs cricket. The latest version of the limited over cricket is more shorter where each team is allowed to bat and bowl for maximum 20 overs. This format of cricket is a popular evening entertainment and the duration of the game or *match* is around two and half hours which is close to other sports such as football, basketball etc. Twenty20 was introduced to create a lively form of cricket which would be attractive to spectators at the ground and viewers on television. The commercial success of this format introduced World Tournament and Indian Premier League (IPL).

The cricket team is a group 11 players consisting of batsmen, bowler and all-rounders. The team should be balanced and diversified to enhance the probability of the success. In addition, the success can also depend on the type of pitch, winning of toss, and sequence of batting or bowling. In general, cricket enthusiasts and experts are of the opinion that batting dominates over bowling. Although, this opinion is widely expressed by many in the cricketing sport, research evidence to support this notion in a formal study is not available in the literature. In this paper, there is a systematic attempt

to support this notion by the application relevant statistical technique such as factor analysis in T20 cricket.

**2. Literature review**

In the literature, Kimber (1993) discussed a graphical method to compare performance of bowlers. Van Staden (2009) proposed a graphical method for comparison of cricketers’ bowling and batting performances. This study can be used to identify different types of players such as offensive batsmen, bowling all-rounders and other important types. Graphical methods are simplistic in nature and give limited insight for detailed comparison. Dey *et al.* (2011) has suggested a multiple decision making approach for evaluation of bowlers performance in Indian Premier League. Kimber and Hansford (1993) also explored statistical analysis of batting in cricket. This study was further extended by Barr and Kantor (2004) by suggesting a mathematical method for comparing and selecting batsmen in cricket. Swartz *et al.* (2006) proposed a simulation procedure for optimal batting order in one day cricket. This work was further extended by Swartz *et al.* (2009) by modeling and simulation for one day cricket. For other dimensions of cricket such as team selection, some advanced mathematical techniques including integer programming and data envelopment analysis have been employed in the literature (see Sharp *et al.*, 2011, Lemmer, 2011, Amin and Sharma, 2012 and Sharma *et al.*, 2012). An advanced statistical technique such as factor analysis has not been employed in the literature of cricket so far whereas it finds frequent application in other research areas. For instance, Valadkhani *et al.* (2008) used factor analysis approach for international portfolio diversification. The limited literature on other dimensions of cricket indicates a strong need to enhance cricketing literature. This paper is perhaps first attempt to apply factor analysis in cricket and contributes substantially to the application of sophisticated methods in cricketing literature.

The remaining part of this paper is organized as follows. Section 2 describes empirical methodology employed in this paper. Section 3 describes data with introduction of IPL 5 along with explanation of batting and bowling capabilities. Section 4 explains the empirical results obtained in this study followed by a section with concluding remarks.

**3. Empirical methodology**

Factor analysis is one of the widely used methods multivariate data analysis (Hair *et al.*, 2007). The purpose of factor analysis is to obtain a reduced set of uncorrelated latent variables using a set of linear combinations of the original variables to maximize the variance of these components. The model can be expressed as follows for a given multivariate set of k variables

$$\begin{aligned}
 r_1 - \mu_1 &= l_{11}f_1 + l_{12}f_2 + \dots + l_{1m}f_m + \varepsilon_1 \\
 r_2 - \mu_2 &= l_{21}f_1 + l_{22}f_2 + \dots + l_{2m}f_m + \varepsilon_2 \\
 &\vdots \\
 &\vdots
 \end{aligned}
 \tag{1}$$

$$r_k - \mu_k = l_{k1}f_1 + l_{k2}f_2 + \dots + l_{km}f_m + \varepsilon_k$$

or in matrix form, this can be written as

$$(\mathbf{r} - \boldsymbol{\mu})_{k \times 1} = \mathbf{L}_{k \times m} \mathbf{F}_{m \times 1} + \boldsymbol{\varepsilon}_{k \times 1}
 \tag{2}$$

with  $m < k$  and where  $\mathbf{r} = (r_1, r_2, \dots, r_k)$  denotes the multivariate vector of various dimensions of batting and bowling,  $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_k)$  is the corresponding mean vector,  $\mathbf{F} = (f_1, f_2, \dots, f_k)$  is a resulting common factor vector,  $\mathbf{L} = [\ell_{ij}]_{k \times m}$  is the matrix of factor loadings,  $\ell_{ij}$  denotes the loading of the  $i$ th variable on the  $j$ th factor and  $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_k)$  is a specific error of  $\mathbf{r}_i$ . Further,  $\mathbf{F}$  and  $\boldsymbol{\varepsilon}$  are independent and their expected values are zero. The covariance of  $\mathbf{F}$  is an identity matrix and covariance of  $\boldsymbol{\varepsilon}$  is  $\hat{\Psi}$ , which is a diagonal matrix.

**2.1 The factor estimation method**

The most widely used method to estimate an orthogonal factor model is principal component analysis (PCA). To use PCA, the data need not be normal and a prior specification of the common factors is not needed. A brief explanation of PCA follows in the next paragraph.

We assume that  $(\hat{\lambda}_1, \hat{e}_1), (\hat{\lambda}_2, \hat{e}_2), \dots, (\hat{\lambda}_k, \hat{e}_k)$  represents a pairs of eigenvalues and eigenvectors of sample covariance matrix  $\hat{\Sigma}_r$ , where  $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \dots \geq \hat{\lambda}_k$  and  $m < k$  means that the number of latent common factors should be less than the number of original variables. The factor loading matrix can be defined as follows

$$\hat{L} = \left[ \begin{matrix} \sqrt{\hat{\lambda}_1} \hat{e}_1 & \dots & \sqrt{\hat{\lambda}_m} \hat{e}_m \end{matrix} \right] \tag{3}$$

The diagonal elements of the matrix  $\hat{\Sigma}_r - \hat{L}(\hat{L})'$  consist of the estimated specific

variances. This means that  $\hat{\Psi} = \text{diag}\{\hat{\Psi}_1, \hat{\Psi}_2 \dots \hat{\Psi}_k\}$ , where  $\hat{\Psi}_i =$

$$s_{ii} - \sum_{j=1}^m \ell_{ij}^2, .$$

The communalities can be estimated using formula  $\hat{h}_i^2 = \ell_{i1}^2 + \ell_{i2}^2$

$+\dots + \ell_{im}^2$ . The residual matrix  $\hat{\Sigma}_r - (\hat{L}(\hat{L})' + \hat{\Psi})$  which is a resultant from the

approximation of  $\hat{\Sigma}_r$  by the principal component solution. The diagonal elements are zero, and if the other elements are also small, we may subjectively take the  $m$  factor model to be appropriate. For detailed information on this method (see Johnson and Wichern, 2002).

**2.2 Factor rotation**

An orthogonal transformation of the factor loadings, as well as the implied orthogonal transformation of the factors, is called *factor rotation* (see Johnson and Wichern, 2002). If  $\hat{L}$  is the  $k \times m$  matrix of estimated factor loadings obtained by principal component analysis, then  $\hat{L}^* = \hat{L}T$ , is a  $k \times m$  matrix of rotated loadings. where  $TT' = T'T = I$  and T is a  $m \times m$  orthogonal matrix. Moreover, the estimated covariance matrix does not change because

$$\hat{L}(\hat{L})' + \hat{\Psi} = \hat{L}T T' \hat{L} + \hat{\Psi} = \hat{L}^*(\hat{L}^*)' + \hat{\Psi}$$

Hence the specific variances  $\hat{\Psi}_i$  and communalities  $\hat{h}_i^2$  does not change.

In practice there are many ways of rotation common factors but the most commonly used method for rotation in literature is the varimax method. This method was proposed by Kaiser ( 1958). Let the rotated matrix of factor loadings be  $L^* = [\ell_{ij}^*]$  and the  $i^{th}$  communalities are given by  $h_i^2$ . We define  $\tilde{\ell}_{ij}^* = \ell_{ij}^*/\hat{h}_i$  to be the rotated coefficients scaled by the square root of the communalities. Then the varimax rotation method chooses the orthogonal transformation matrix T which makes V as maximum as possible where

$$V = \frac{1}{k} \sum_{j=1}^m \left[ \sum_{i=1}^k (\tilde{\ell}_{ij}^*)^4 - \frac{1}{k} \left( \sum_{i=1}^k \tilde{\ell}_{ij}^{*2} \right)^2 \right]$$

When V is maximum it means that the squares of the loadings on each factor are spread out as much as possible. The purpose is to interpret the common factors by searching groups of very large and very small coefficients in any column of the rotated matrix of factor loadings.

**4. Data**

To illustrate the proposed model of factor analysis, we have taken data from IPL session 5 in 2012. A set of 85 batsmen and 85 bowler’s data were selected for this study. The data files have not been reported in this paper because of size of files but they are available from the author upon request. In this section, we begin with a brief outline about IPL5 followed by explanation about various player’s capabilities.

**4.1 IPL 5 (Indian Premier League session 5)**

The **Indian Premier League (IPL)** is a professional league initiated by Board of Control for Cricket in India (BCCI) for Twenty20 cricket championship in India. IPL5-T20 was concluded in the month of April-May 2012 and was a huge economic and entertainment forum. Spectators on the field and viewers of the television were entertained to the fullest. In this fifth session of IPL, there were a total of 9 teams on the names of famous cities of India like Chennai Super Kings, Royal Challengers of Bangalore, etc .These teams choose players via an auction where maximum players are chosen from the list of players playing in India in different formats of the game whereas

some players are chosen from different cricket playing countries. The maximum number of foreign players can be taken into a team is four. The cumulative reach for 74 IPL 5 matches was recorded at 163 million against 162 million for 73 matches in IPL 4. The brand value of IPL 5 was estimated to be around US\$2.99 billion.

#### 4.2 Batting statistics

In this study, we consider five important measures of batting statistics such as highest individual score (HS), average batting performance, strike rate (SR), numbers of fours (4s), and number of sixes (6s). **HS:** For a batsman, the highest individual score (HS) is the maximum number of runs scored by a batsman in one match during a tournament. **Average:** The average batting performance is expressed by  $R/m$  where  $R$  denotes the number of runs scored and  $m$  the number of times the batsman was out. **SR:** The batting strike rate can be expressed using the ratio  $R/b$  where  $R$  denotes the number of runs scored and  $b$  denotes the number of balls faced by a player. **4s:** The number of 4s hit by the batsman. **6s:** The number of 6s hit by the batsman.

#### 4.3 Bowling statistics

Sharp *et al.* (2011) and Lemmer (2011) used three bowling measures such as bowler's economy rate, bowling average, and bowling strike rate. Bowler's economy rate can be expressed by  $TR/O$  where  $TR$  is the total number of runs conceded by a bowler and  $O$  is the total number of overs bowled by a bowler. Bowling average can be expressed as  $TR/W$  where  $TR$  is the total runs conceded by a bowler and  $W$  is the total number of wickets taken by a bowler. Bowling strike rate can be expressed as  $TB/W$  where  $TB$  is the total number of balls bowled by a bowler and  $W$  is the total number of wickets taken by a bowler.

### 5. Empirical results

The factor analysis as explained was performed using PCA to explain items validity as well as groups of items into meaningful clusters. As far as the suitable rotation policy is concerned, an orthogonal varimax rotation was used since it assists in optimizing the number of variables. The higher loadings on a particular factor result in identification of each variable with a single factor. The factor loadings are given in table I. We have considered only those items whose factor loading are higher than 0.5. In selecting the extraction of factorial groups, the Kaiser criterion was adopted. As per the Kaiser criterion, all factors should be accepted whose eigenvalues are higher than 1.0. Therefore, two factors were extracted (see Figure 1).

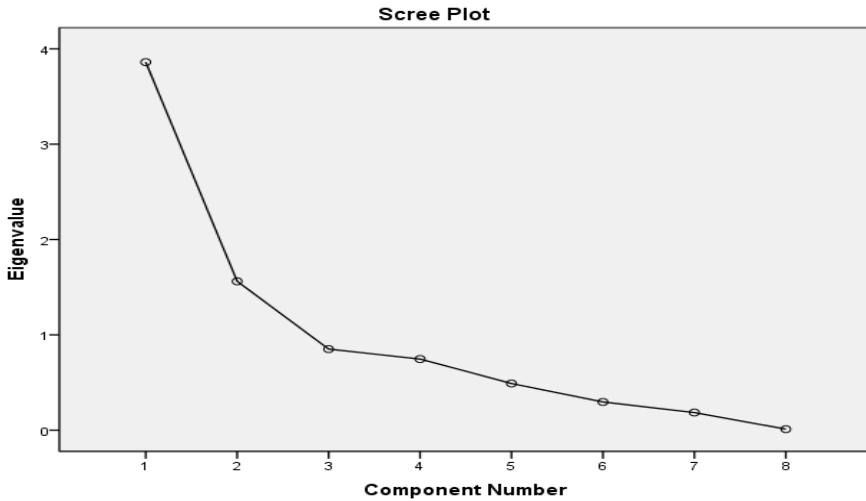


Figure 1: Scree plot

The variance explained by first factor (Batting) was 48.51 per cent and the variance explained by second factor (Bowling) was 20.23 per cent. These two extracted factors accounted for 68.74 per cent of the total variance explained in this research and are shown in Table I. The variance explained by batting is much higher than the variance explained by bowling, which explains the higher importance of batting as compared to bowling in T20 cricket.

Table I: Rotated Component Matrix

Variables	Rotated factor loadings and communalities		
	Factor 1	Factor 2	Communalities
Highest Score by batsman	<b>0.768</b>	-0.377	0.732
Average of a batsman	<b>0.767</b>	-0.211	0.632
Strike rate of a batsman	<b>0.770</b>	0.106	0.604
Number of fours hit by a batsman	<b>0.641</b>	-0.409	0.578
Number of sixes hit by a batsman	<b>0.876</b>	-0.174	0.798
Bowling average	-0.118	<b>0.969</b>	0.952
Bowling economy	-0.201	<b>0.538</b>	0.330
Bowling strike rate	-0.101	<b>0.929</b>	0.873
Percentage of variance	<b>48.51</b>	<b>20.23</b>	
Cumulative percentage	<b>48.51</b>	<b>68.74</b>	

## 6. Concluding remarks

The need for incorporation of advanced research methods has increased in the past few years due to the huge investments and commercial importance of sports. Whereas, many professional sports have enjoyed the attention of well-established research methods, cricket still continues to be a sport that has not received adequate attention. Therefore, this paper is an important step in the right direction. The paper has employed factor analysis for the first time in cricketing game research problem. The factor analysis has been employed to explore the interrelationship among various dimensions of T20 cricket. Data of 85 batsmen and 85 bowlers with various dimensions of batting and bowling was used. The five dimensions have been grouped into factor one (batting), whereas three dimensions have been grouped into factor two (bowling). The variance explained by factor one (batting) is much higher than factor two (bowling) which shows clear dominance of batting capability over bowling capability. This was an important aspect of cricket often used by team selectors; the results from this paper now provides more support give higher priority to batting capability over bowling capability. The method of factor analysis which has been successfully used to explain an important hypothesis in cricket in this paper may similarly be employed in other sports and constitutes the scope for future research.

## Acknowledgement

Author would like to thank the referees and Editorial Board of the Journal of Reliability and Statistical Studies for their valuable suggestions in improving the quality of this paper.

## References

1. Barr, G.D.I. and Kantor, B.S. (2004). A criterion for comparing and selecting batsmen in limited overs cricket, *Journal of the Operational Research Society*, 55, p. 1266-1274.
2. Gholam, R. A. and Sharma, S.K. (2012). Cricket team selection using data envelopment analysis, *European Journal of Sport Science*, DOI:10.1080/17461391.2012.705333.
3. Hair, J.F., Black, W.C., Babin, Anderson, R.E. and Tatham, R.L. (2007). *Multivariate Data Analysis*, 6th ed., Prentice-Hall, Upper Saddle River, NJ.
4. Johnson, R.A. and Wichern, D.W. (2002). *Applied Multivariate Statistical Analysis*, 5th ed., Prentice-Hall, Upper Saddle River, NJ.
5. Kimber, A.C. (1993). A graphical display for comparing bowlers in cricket, *Teaching Statistics*, 15, p. 84-86.
6. Kimber, A.C. and Hansford, A.R. (1993). A statistical analysis of batting in cricket, *Journal of the Royal Statistical Society, Series A* 156, p. 443-455.
7. Lemmer, H.H. (2011). Team selection after a short cricket series, *European Journal of Sport Science*, DOI: 10.1080/17461391.2011.587895.
8. Preston, I. and Thomas, J. (2000). Batting strategy in limited overs cricket, *Statistician*, 49(1), p. 95–106.
9. Dey, P. K., Ghosh, D. N., and Mondal, A. C. (2011). A MCDM Approach for Evaluating Bowlers Performance in IPL, *Journal of Emerging Trends in Computing and Information Sciences*, 2(11).
10. Sharma, S. K., Amin, G. R. and Gattoufi, S. (2012). Choosing the best Twenty20 cricket batsmen using ordered weighted averaging, *International Journal of Performance Analysis in Sports*, 12(3), p. 614-628.

11. Sharp, G.D., Brettenny, W.J., Gonsalves, J.W., Lourens, M. and Stretch, R.A. (2011). Integer optimization for the selection of a Twenty20 cricket team, *Journal of the Operational Research Society*, 62, p. 1688-1694.
12. Swartz, T.B., Gill, P.S., Beaudoin, D. and de Silva, B.M. (2006). Optimal batting orders in one-day cricket, *Computers and Operations Research* 33, p. 1939- 1950.
13. Swartz, T.B., Gill, P.S. and Muthukumarana, S. (2009). Modelling and simulation for one-day cricket, *The Canadian Journal of Statistics*, 37, p. 143-160.
14. Valadkhani, A., Chancharat, S. and Harvie, C. (2008). A factor analysis of international portfolio diversification, *Studies in Economics and Finance*, 25(3), p. 165- 174.
15. Van Staden, P. J. (2009). Comparison of cricketers' bowling and batting performances using graphical displays, *Current Science*, 96(6), p. 764–766.