# An Introductory Application of Principal Components to Cricket Data

Ananda B. W. Manage
Stephen M. Scariano
Sam Houston State University

*Journal of Statistics Education* Volume 21, Number 3 (2013),
www.amstat.org/publications/jse/v21n3/scariano.pdf



---

**Key Words:** Multivariate statistics; Sports data; Principal components; Cricket.

## Abstract

Principal Component Analysis is widely used in applied multivariate data analysis, and this article shows how to motivate student interest in this topic using cricket sports data. Here, principal component analysis is successfully used to rank the cricket batsmen and bowlers who played in the 2012 Indian Premier League (IPL) competition. In particular, the first principal component is seen to explain a substantial portion of the variation in a linear combination of some commonly used measures of cricket prowess. This application provides an excellent, elementary introduction to the topic of principal component analysis.

## 1. Introduction

The goal of this paper is to demonstrate, at an elementary level, the utility of principal component analysis in sports data. In particular, we discuss the applicability of principal components in ranking cricket players. For those new to the game, we begin with a brief introduction, but a complete description of the game rules and regulations can be found on the web at www.cricket-rules.com. Cricket is one of the first sports to use statistics as a tool for illustration, comparison and prediction (Bennett 1998, pp. 83-103). Although cricket, like baseball, soccer, and basketball, is a sport replete with statistical applications, much less statistical work has been done for cricket in comparison to these other sports.

Like baseball and soccer, cricket is an open-field game played by two adversarial teams, and Figure 1 shows a typical field inside a large stadium. There are basically three types of cricket

games played today that enjoy a wide fan base: test cricket, One Day International (ODI) and Twenty20. The latter format, although new, is rapidly becoming popular among cricket fans.

**Figure 1**. A Cricket Field

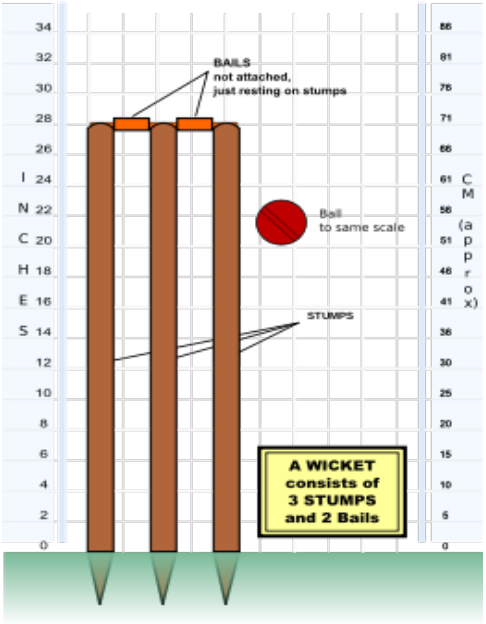
## 2. Cricket Game Description

Before play commences, the two team captains call the toss of a coin, and the winning captain decides which team shall *bat* or *bowl* first. The captain who wins the toss makes his decision on the basis of tactical considerations regarding the strengths/weaknesses of the competing teams as well as field conditions. In One Day International (ODI) matches, *50-overs* (300 = *50 x* 6 legal (game-permissible) ball deliveries) are allowed for each side, while *20-overs* (120 = *20 x* 6 legal ball deliveries) are allowed in Twenty20 matches.

**Figure 2a**. Cricket Pitch

There are 11 players on each team. *Batsmen* play in pairs. *Bowlers* (similar to baseball pitchers) are not allowed to throw the ball; instead, they must use a "stiff-arm" ball delivery technique. There are two basic types of bowling: fast and spin. A fast bowler mainly uses speed as the tactic while a spinner may spin the ball, making it harder for the batsman to judge, and there are other factors, like swing and bounce, used to distract batsmen. The rectangular area in the middle of the field is called the *pitch*, and this is where most of the action happens. There are two sets of sticks, called *wickets*, which are placed 22 yards apart in the pitch (Figures 1, 2a, 2b).

**Figure 2b**. Wicket (3 Stumps and 2 Bails)




A batman's aim is to *bat* and score *runs* while simultaneously protecting the wickets. The bowler's aim is to deliver (*bowl*) the ball at the wicket so that the ball hits and disrupts the wicket.

This is one of several ways that a batsman can be called *out*. In one turn, a bowler delivers a set of six balls, which is called an *over*. Once a batsman hits the ball sufficiently far, he runs to the (*non-striker*) wicket while the other batsman at the non-striker wicket simultaneously runs to the *striker* wicket, accumulating a single run. Batsman will be called *out* if a fielder breaks the wicket with the ball before the batsman reaches his wicket. Each switch like this gives one run to the striker, but the players may continue to switch as long it is safe to do so. There is an outside boundary around the cricket field. When a batsman hits the ball and it reaches the boundary after hitting the ground inside the boundary, four runs (a *boundary*) are awarded to the batsman. However, if the ball goes over the boundary without hitting the ground, six runs are awarded to the batsman.

There are several ways a batsman can be *dismissed*. If a batsman misses the ball and it hits the wicket, the batsman is out, and this dismissal is called *bowled*. If a fielder catches the ball in full (without it first hitting the ground) after it was hit by the batsman, he is out and this dismissal is called *caught*. If the ball hits the body (usually a leg) of the batsman, and if the umpire decides the ball would have hit the wicket otherwise, the batsman is called out and this dismissal is called

*LBW* (Leg Before Wicket). If a fielder breaks the wicket using the ball before a batsman reaches his end while attempting to score, the batsman is called out and this dismissal is called a *run out*. If the wicket keeper, the fielder who stays behind the wicket at the striker end (similar to a catcher in baseball), breaks the wicket with the ball while the striker batsman is outside his safe zone, the *crease*, the batsman is out and this dismissal is called *stumped*. If a batsman breaks the wicket with the bat or his body while attempting to score, he is called out and this dismissal is called *hit wicket*. There are a few other methods of dismissing batsman which are not so common, such as *handle the ball* (touching the ball while it is live), *obstructing the fielder* (deliberately obstructing a fielder), *hit the ball twice* (hitting the ball twice for any reason other than to defend his wicket from being broken by the ball) and *timed out* (new batsman taking more than two minutes to come to bat after a wicket has fallen). This concludes our brief introduction to the basic rules of cricket, but much more detailed information can be found at www.cricket-rules.com.

There are several studies in the literature related to performance analysis in cricket. Barr and Kantor (2004) proposed a method based on batting averages and strike rates. Lewis (2008) analyzed player performance using Duckworth/Lewis percentage values. Borooah and Mangan (2010) explored batting performance for test matches. Van Staden (2009) used a graphical method to analyze batting and bowling performance in cricket. The series of papers Lemmer (2004), Lemmer (2008b) and Lemmer (2012) considered performance analysis using averages and strike rates for bowling and batting. Lakkaraju and Sethi (2012) described a Sabermetrics-style principle to the game of cricket to analyze batting performance. The method we describe in the next section uses several other factors in addition to averages and strike rates.

## 3. Cricket Variables Influential on Batting and Bowling Performance

Indian Premier League (IPL) is a professional Twenty20 championship cricket league in India that has become very popular among cricket fans worldwide. Twenty20 is the latest format of cricket, and a typical match lasts about 3.5 hours. When compared to One Day International (ODI) or Test Cricket games, its fast-paced style and shortened duration are two key reasons for the increasing popularity of Twenty20 in recent years. Beginning in 2008, IPL had completed its fifth consecutive season by May 2012. For the 2012 competition there were 9 competing teams: Royal Challengers Bangalore, Rajasthan Royals, Pune Warriors India, Mumbai Indians, Kolkata Knight Riders, Kings XI Punjab, Delhi Daredevils, Deccan Chargers, and Chennai Super Kings.

There are several ways that a franchise acquires players. One of these methods is to buy players at player auctions. Obviously, franchises pay higher salaries for quality players based on their performance. At a tournament completion, it is always interesting to ask, "How did these quality players actually perform?" Performance measures can be used to decide a particular player's price in future auctions, and there are several indicators typically used to measure the performances of players.

When considering batsmen, the goal in limited-overs cricket, like Twenty20, is to score as many as runs as possible using as few balls as possible. On the other hand, building longer innings instead of trying to score runs for each ball is the key goal in test cricket. For a batting analysis, there is set of widely-recognized variables that can be used to measure the quality of each

batsman. These variables are commonly used by cricket commentators and sports authorities, and are also shown in scoreboards to describe player profiles.
(cf: http://www.espncricinfo.com/ci/content/player/49535.html).

For convenience, the batting and bowling data (with codebook) used in this study are www.amstat.org/publications/jse/v21n3/scariano/batting_data.csv; www.amstat.org/publications/jse/v21n3/scariano/bowling_data.csv; and www.amstat.org/publications/jse/v21n3/scariano/code_book.txt.

The variables used here are:

**Runs:** The total number of runs scored by a player in the IPL 2012 season. Higher values indicate stronger performance.

**Batting Average** (**Ave**): The total number of runs a batsman has scored divided by the total number of times he has been called out in the IPL 2012 season. Higher values indicate stronger performance. However, for a batsman with several "not out" cases, this number overrates the batsman, which is a weakness in this measure, and this is why it should not be used as the only variable for batting performance analysis.

**Batting Strike Rate** (**SR**): The Batting Strike Rate is defined as the number of runs scored per 100 balls faced by a batsman in the IPL 2012 season. Again, higher values indicate stronger performance. An aggressive batting style is always helpful in shorter versions of limited-overs cricket matches like Twenty20. However, a high strike rate accompanying a low batting average is not desirable.

**Fours**: The total number of boundaries (fours = four runs) made in the IPL 2012 season by a batsman. Higher values indicate stronger performance. Scoring boundaries is a great way to increase the number of runs without wasting resources, and it helps increase the batting average and strike rate.

**Sixes**: The total number of sixes (= six runs) made in the IPL 2012 season by a batsman. As before, higher values indicate stronger performance. Scoring sixes usually enhances team scoring momentum and can also diminish the momentum of the opposing bowler.

In addition to these variables and after careful consideration, we constructed a new variable which combines the number of Centuries (100 or more runs in an innings) together with the number of Fifties (50 or more runs in an innings but less than 100 runs in an innings). Because there were not many centuries in IPL 2012, we created the variable, HF (= half-centuries). So,

**HF** = (2 *x* Number of Centuries) + Number of Fifties. Higher values are indicative of exceptional performance, and it is always advantageous to build partnerships and play longer innings in any cricket format. There were only six centuries for the IPL 2012 tournament.

This collection of variables can be used to study batting performance via Principal Components Analysis. Other variables, like batting position (batting order), winning the toss, and the so-

called “home-field advantage” might also be considered, but incorporating those here would be beyond the scope and objective of this article.

In what follows, we use the six “batting variables”: **Runs**, **Ave**, **SR**, **Fours**, **Sixes** and **HF** defined above to conduct our analysis. As expected, some of these variables are likely to be highly related (i.e, correlated) to each other, making it difficult to construct an overall batting-performance measure, which is our goal.

Figure 3 shows individual histograms for our “batting variables” based on the performance of the ninety batsmen who played at least five innings in the 2012 IPL season. Figure 4 depicts a matrix plot, and we see some significant correlations between these variables. For example, the plot shows that **Runs** and **Fours**, **Runs** and **Sixes**, and **Ave** and **Runs** are considerably correlated, as might be anticipated. This shows the necessity of using a technique which is

**Figure 3**. Individual Histograms for Batting Variables

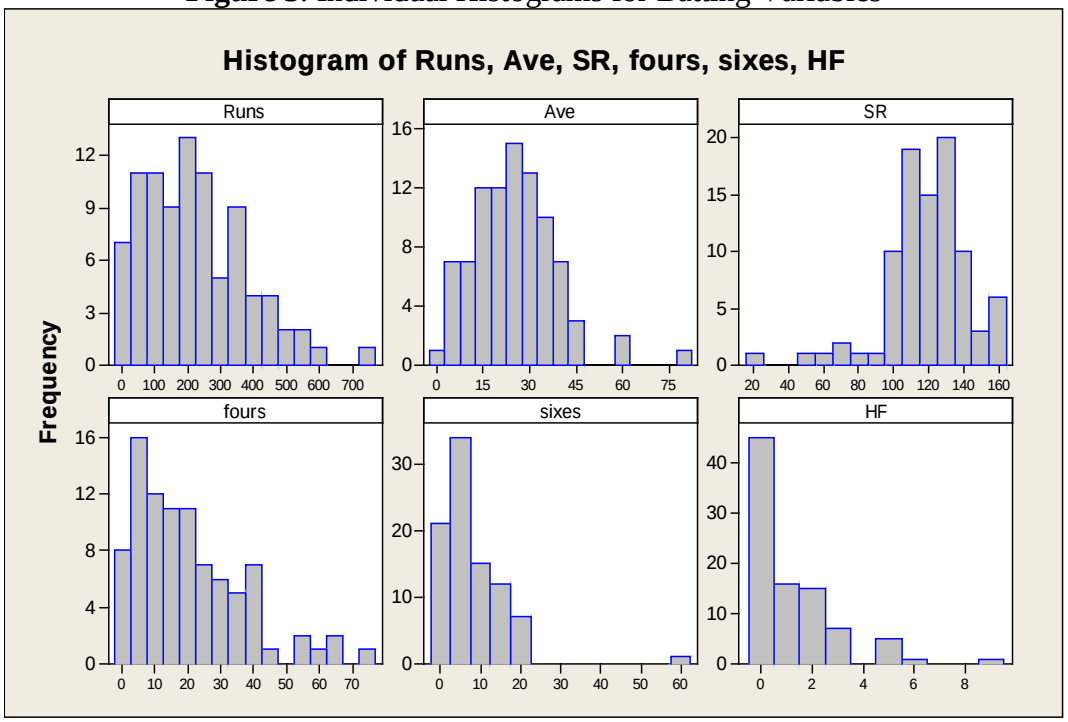
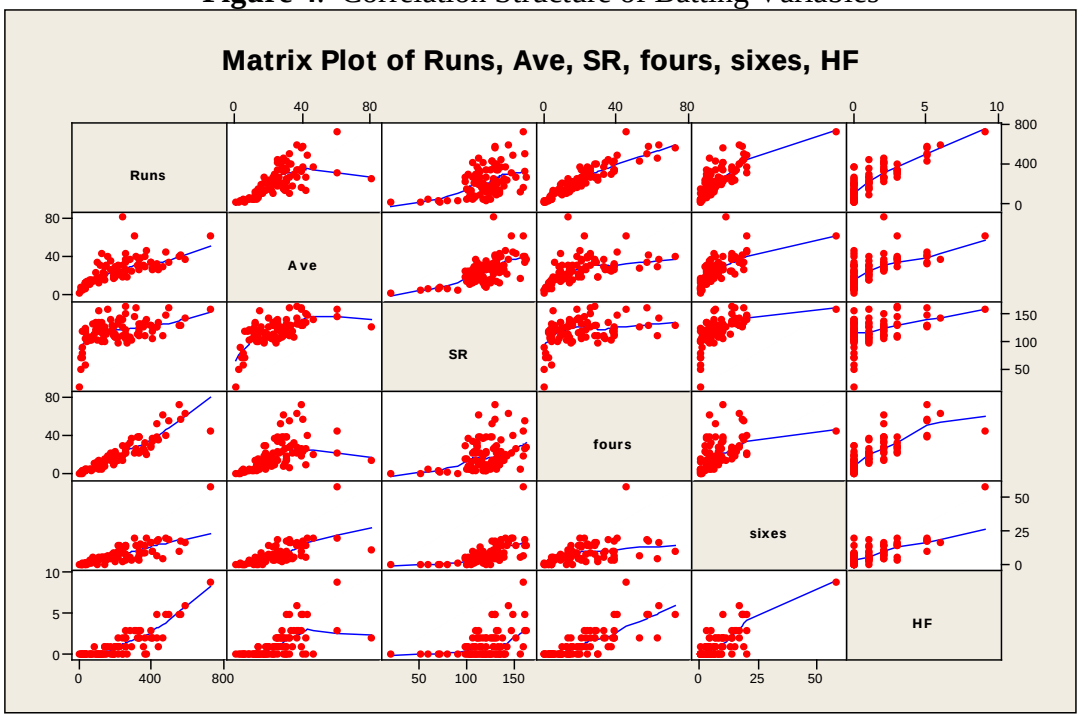


**Figure 4**. Correlation Structure of Batting Variables




capable of handling correlated data in any reasonable attempt to study batting performance. As stated earlier, high values for each of these variables indicates better batting performance in a univariate sense, and each one measures a different quality of a batsman. But, the primary concern is their joint contribution to batting performance in a multivariate sense. Constructing an overall measure of batting performance by collapsing these correlated variables is a key goal of this paper. Of course, a weighted-average type measure would be ideal, and this is what Principal Component Analysis offers in this context.

Turning now to bowling performance, we use a set of widely-recognized variables to measure the quality of each bowler. These variables are the ones widely used by cricket announcers to describe player performance and are likewise shown in scoreboards to describe player profiles (http://www.espncricinfo.com/ci/content/player/49535.html).

**Wickets** (**Wkts**): The number of wickets taken by a bowler. There are ten possible wickets for an innings and there should be at least five bowlers, each of whom can bowl a maximum of four *overs*. A bowler's goal is to take the maximum number of wickets from the *overs* that he bowls, so taking a large number of wickets from batsmen is one performance measure for bowlers. However, like the total number of runs statistic for a batsman, the number of wickets taken is not sufficient to measure the quality of a bowler. The goal of a bowler is to get the maximum number of wickets by using a minimum number of balls while simultaneously conceding a minimum number of runs.

**Bowling Average** (**Ave = Runs/Wkts**): The average number of runs conceded per wicket. Here, lower values are preferred since a bowler's goal is to concede the minimum number of runs while simultaneously earning the maximum number of wickets.

**Strike Rate** (**SR = Balls/Wkts**): The average number of balls bowled per wicket taken. Lower values are preferred since a bowler should try to bowl the minimum number of balls per wicket.

**Economy Rate** (**Econ** = **Runs/(*overs* bowled)**: The average number of runs conceded per *over*. Lower values are preferred since this is the run-rate against a specific bowler for a batting team. Therefore, the bowler's aim is to keep this measure as small as possible.

As with batsmen, other variables, like the significance of the wickets taken (wickets taken at the front end of the batting order are usually harder to get than those towards the end) might also be considered.

Figure 5 shows individual histograms for our "bowling variables" based on the performance of the eighty-three bowlers who bowled at least ten *overs* in the 2012 IPL season.

**Figure 5**. Individual Histograms for Bowling Variables

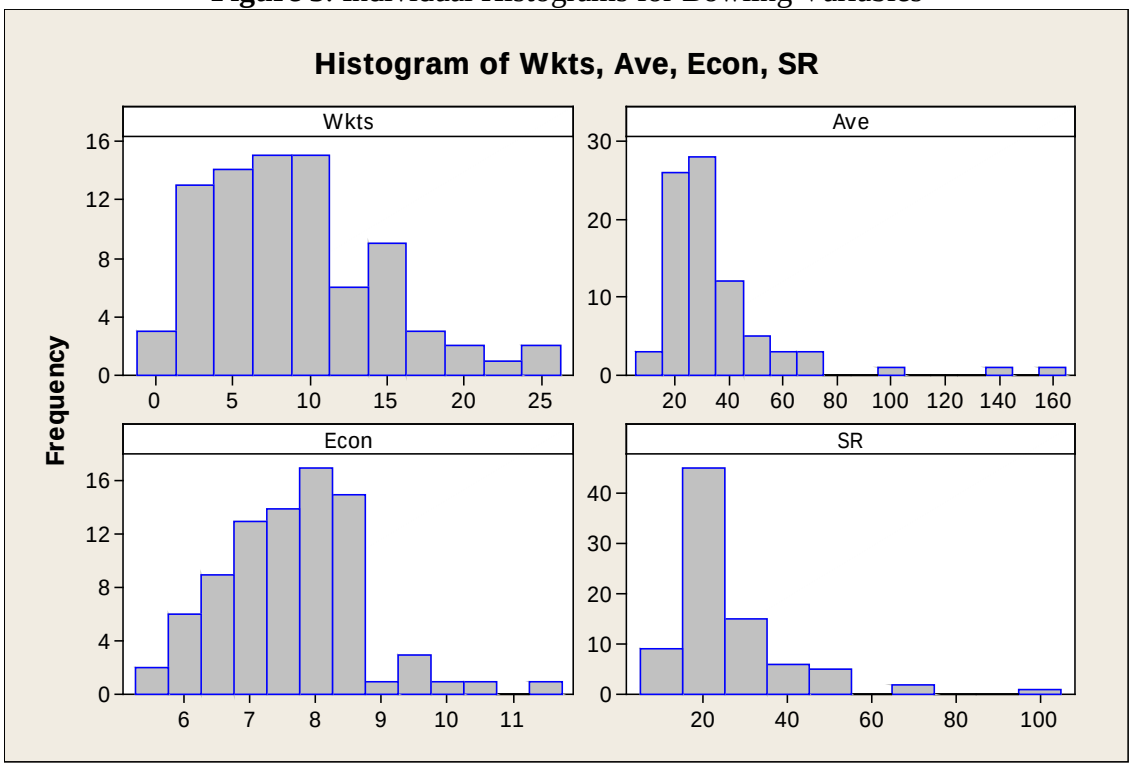
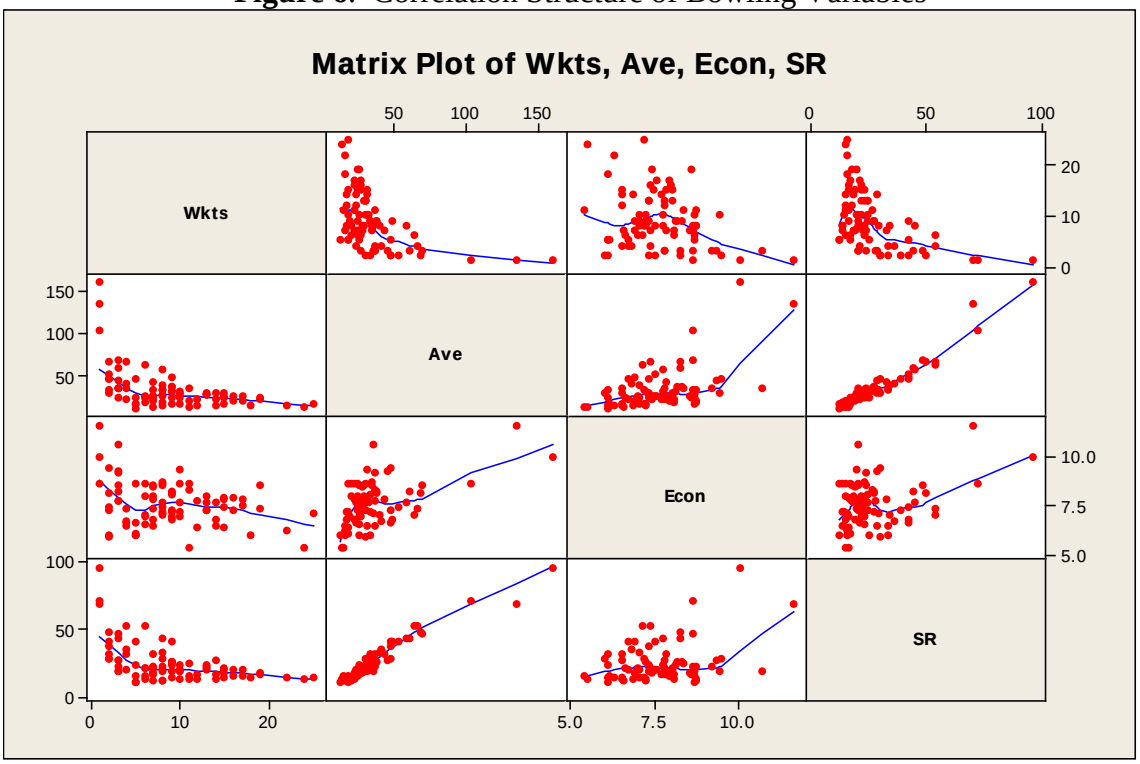


**Figure 6**. Correlation Structure of Bowling Variables




The correlation structure of the four "bowling variables" which we use is shown in the matrix plot of Figure 6. As we see, **Bowling Average** and the **Strike Rate** are highly positively correlated. All the other variables are somewhat negatively correlated with the number of wickets. However, each one of these variables measures a different quality of a bowler, even though they are correlated. Constructing an overall measure of performance by using some kind of weighted averaging would be the ideal way to handle this situation. This suggests the potential usefulness of the Principal Component Analysis technique for bowler performance as well.

## 4. Principal Components and Cricket Performance

Principal Component Analysis (PCA) is a nonparametric variable reduction technique well-suited for correlated data that can be effectively used in our context. One objective of principal component analysis is to collapse a set of correlated variables into fewer uncorrelated variables as linear combinations of the original variables. Readers can find excellent introductions to Principal Component Analysis in the works of Johnson and Wichern (2007), Dawkins (1989), and Watnik and Levine (2001). We will give a brief introduction to the PCA technique.

PCA is particularly useful when data on a number of useful variables has been gathered, and it is plausible that there is some redundancy in those variables. Here, redundancy is taken to mean that our cricket performance variables are correlated with one another because, in some unknown sense, they might be measuring similar player-performance attributes. PCA aims to reduce the observed variables down to a smaller number of principal components, sometimes called auxiliary variables (optimized linear combinations of the original variables), which account for

most of the variation occurring in the originally observed variables. These can be utilized to provide summarized measures of performance.

Briefly, given a random vector $\boldsymbol{X} = (X_1, X_2, \ldots\ldots, X_p)^t$ consisting of $p$ random variables, having covariance matrix $\boldsymbol{\Sigma}$ and eigenvalue-eigenvector pairs $(\lambda_1, \mathbf{e}_1)$, $(\lambda_2, \mathbf{e}_2)$ , ……, $(\lambda_p, \mathbf{e}_p)$, where $\lambda_1 \geq \lambda_2 \geq \ldots\ldots \geq \lambda_p \geq 0,$ the ***i*th *principal component***, say $L_i$, is defined as $L_i = \mathbf{e}_i^t X = e_{i1}X_1 + e_{i2}X_2 + \ldots\ldots\ldots + e_{ip}X_p$ for $i$ = 1,2,……, $p$ where $(e_{i1}, e_{i2}, \ldots\ldots\ldots, e_{ip})$ are the components of eigenvector $\mathbf{e}_i^t$. From this, it can be seen that the principal components are linear combinations of the original random variables of interest. Moreover, it can be shown that (i) if $Y_i = a_i^t X = a_{i1}X_1 + a_{i2}X_2 + \ldots\ldots\ldots + a_{ip}X_p$ is any other linear combination of these original variables, then for the first principal component, $Var(L_1) = \lambda_1 \geq Var(Y_i),$ (ii) $Cov(L_i, L_j) = 0,$ for $i \neq j,$ and (iii) $\sum_{i=1}^{p} Var(X_i) = \sum_{i=}^{p} Var(L_i).$ Property (i) gives us hope that the principal components $L_i$ can be used to capture the important signals aggregately contained in the original variables $X_1, X_2, \ldots\ldots, X_p$ , while property (ii) shows that this can be done without redundancy. Property (iii) provides a means of identifying the contribution of each principal component to signal detection. To see this, notice that

$$\begin{aligned} \text{Total Variance } &= Var(X_1) + Var(X_2) + \ldots\ldots + Var(X_p) \\ &= Var(L_1) + Var(L_2) + \ldots\ldots + Var(L_p) \\ &= \lambda_1 + \lambda_2 + \ldots\ldots + \lambda_p. \end{aligned}$$

So, the proportion of the total population variance due to the $j^{\text{th}}$ principal component is the ratio $\lambda_j \times \left( \sum_{i=1}^{p} \lambda_i \right)^{-1}$. If the first few principal components capture a large percentage of the total variance, then it is plausible that these new variables can be used in place of the original variables without much loss of information (signal). Customarily, variables measured on different scales should first be standardized before commencing PCA. If standardization is not performed, the resulting principle components will be dominated by the variables with maximum variance, which does not meet the goal of an overall performance measure. The nature of 2012 IPL cricket data requires this standardization prior to PCA, and we have done so.

## 4.1 Ranking Batsmen using the First Principal Component

This analysis includes the **Runs, Batting Average** (**Ave**), **Batting Strike Rate** (**SR**), **Fours**, **Sixes**, and **HF** variables, discussed in section 3, for all batsmen who have played at least five innings in the 2012 IPL season. This accounts for 90 total batsmen (Table 6). The number of innings (always plural) played for a batsman is the number of games in which he actually bats; however, in limited-overs cricket, the game could conclude before a batsman ever gets to bat, which would not count as an innings for that particular batsmen.

**Table 1.** Sample Correlation Matrix for 90 Batsmen.

| | **Runs** | **Ave** | **SR** | **Fours** | **Sixes** | **HF** |
|---|---|---|---|---|---|---|
| **Runs** | **1.00** | 0.69 | 0.49 | 0.92 | 0.77 | 0.84 |
| **Ave** | 0.69 | **1.00** | 0.62 | 0.55 | 0.68 | 0.62 |
| **SR** | 0.49 | 0.62 | **1.00** | 0.38 | 0.58 | 0.43 |
| **Fours** | 0.92 | 0.55 | 0.38 | **1.00** | 0.52 | 0.78 |
| **Sixes** | 0.77 | 0.68 | 0.58 | 0.52 | **1.00** | 0.77 |
| **HF** | 0.84 | 0.62 | 0.43 | 0.78 | 0.77 | **1.00** |

Values for each of these variables were collected together into a $(6 \times 1)^t$ column vector of the form (**Runs**, **Ave**, **SR**, **Fours**, **Sixes**, **HF**)$^t$ for each of 90 batsman. These we call the *batting vectors*. Once data have been obtained, the $(6 \times 6)$ sample correlation matrix associated with the sample batting vectors may be examined for the correlation structure inherent in these cricket variables.

Because these variables are measured on very different scales, they must be standardized before PCA analysis. However, the process of finding the principal components by using the standardized variables is equivalent to finding principal components by using the correlation matrix instead of the covariance matrix. All modern statistical software packages utilize programs for computing the eigenvalue-eigenvector pairs of a sample correlation matrix. Here, we use Minitab® to accomplish this task. Table 1 shows the sample correlation matrix for the ninety batting vectors.

Table 2 gives the ordered eigenvalues and percentage of total variability attributed to each, while Table 3 shows the eigenvector coefficients for all six principal component (here, PC1 – PC6), which are listed only for completeness. The eigenvalue-eigenvector pair for the first principal component is highlighted in Table 3.

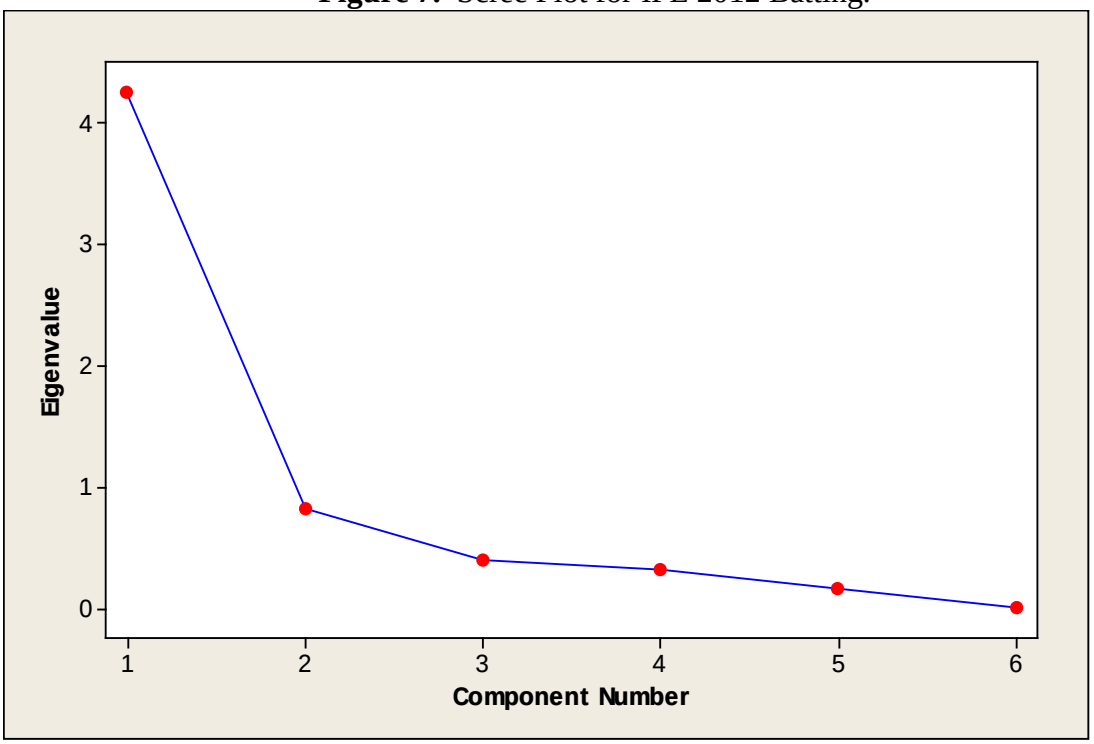
**Table 2.** Ordered eigenvalues and corresponding Percentages of Total Variability for Batsmen.

| Eigenvalue | **4.255** | **0.827** | **0.412** | **0.326** | **0.164** | **0.017** |
|---|---|---|---|---|---|---|
| Total Variability | 70.91% | 13.79% | 6.87% | 5.43% | 2.73% | 0.28% |

**Table 3.** Eigenvalue and Eigenvector Pairs for the Sample Correlation Matrix in Table 1.

| Eigenvalues: | **4.255** | 0.827 | 0.412 | 0.326 | 0.164 | 0.017 |
|---|---|---|---|---|---|---|
| Variable | **PC1** | PC2 | PC3 | PC4 | PC5 | PC6 |
| **Runs** | **0.458** | -0.266 | -0.110 | 0.005 | -0.458 | 0.705 |
| **Ave** | **0.398** | 0.331 | 0.006 | -0.847 | 0.101 | -0.061 |
| **SR** | **0.325** | 0.698 | -0.450 | 0.433 | 0.119 | 0.056 |
| **Fours** | **0.406** | -0.474 | -0.508 | 0.033 | -0.097 | -0.585 |
| **Sixes** | **0.417** | 0.179 | 0.669 | 0.249 | -0.395 | -0.358 |
| **HF** | **0.432** | -0.276 | 0.281 | 0.178 | 0.775 | 0.161 |

**Figure 7.** Scree Plot for IPL 2012 Batting.




Consequently, the first principal component for batsmen is

$$L_1 = 0.458*\textbf{Runs} + 0.398*\textbf{Ave} + 0.325*\textbf{SR} + 0.406*\textbf{Fours} + 0.417*\textbf{Sixes} + 0.432*\textbf{HF},$$

where now the variables **Runs**, **Ave**, **SR**, **Fours**, **Sixes**, **HF** are understood to have already been individually standardized. The first entry in the second row of Table 2 shows that almost 71% of the Total Variability can be explained by this first principal component. Moreover, its corresponding eigenvalue 4.255 is the only one which is greater than 1. Kaiser (1960) suggests

retaining all principal components whose corresponding eigenvalues exceed unity. If the ordered eigenvalues are plotted sequentially, then the resulting visual presentation is commonly called a *Scree plot*, which may be used to ascertain the appropriate number of principal components to retain in a particular application. To do this, one looks for an *elbow* (bend) in the plot. The number of useful principal components is then taken to be the abscissa of the point beyond which all remaining eigenvalues add relatively small contributions to the total variability. Figure 7 shows the Scree plot for the IPL 2012 season, suggesting that since the elbow is at abscissa two, it is reasonable to use only the first principal component which explains 71% of the total variability.

Following Johnson and Wichern (2007, p. 452), who discuss a "general stock-market component" in a different context where the first principal component is useful, we refer to our first principal component as the *general-batting-performance-index*, which is a type of weighted average of all six variables used. Here, the coefficients of the first principal component are all positive, so larger values of $L_1$ indicate better player performance. This justifies that we should rank (largest to smallest) the players based on the first principal component.

**Table 4.** Top Ten Batsmen for IPL 2012 (minimum 300 runs) using First Principal Component, $L_1$

| Batsman | Matches | Innings | Runs | Ave | SR | Hundreds | Fifties | Fours | Sixes | HF | $L_1$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| C.H. Gayle | 15 | 14 | 733 | 61.08 | 160.74 | 1 | 7 | 46 | 59 | 9 | 8.47 |
| G. Gambhir | 17 | 17 | 590 | 36.87 | 143.55 | 0 | 6 | 64 | 17 | 6 | 4.59 |
| V. Sehwag | 16 | 16 | 495 | 33.00 | 161.23 | 0 | 5 | 57 | 19 | 5 | 4.12 |
| S. Dhawan | 15 | 15 | 569 | 40.64 | 129.61 | 0 | 5 | 58 | 18 | 5 | 4.10 |
| A.M. Rahane | 16 | 16 | 560 | 40.00 | 129.33 | 1 | 3 | 73 | 10 | 5 | 4.00 |
| C.L. White | 13 | 13 | 479 | 43.54 | 149.68 | 0 | 5 | 41 | 20 | 5 | 3.88 |
| R.G. Sharma | 17 | 16 | 433 | 30.92 | 126.6 | 1 | 3 | 39 | 18 | 5 | 2.90 |
| K.P. Pietersen | 8 | 8 | 305 | 61.00 | 147.34 | 1 | 1 | 22 | 20 | 3 | 2.86 |
| A.B. de Villiers | 16 | 13 | 319 | 39.87 | 161.11 | 0 | 3 | 26 | 15 | 3 | 2.31 |
| F. du Plessis | 13 | 12 | 398 | 33.16 | 130.92 | 0 | 3 | 29 | 17 | 3 | 2.11 |

Table 4 gives the top 10 batsmen who scored a minimum of 300 runs in the IPL 2012 season when ranking using the first principal component, $L_1$. We chose this 300 run limit for purposes of comparison with a ranking method due to M. Ramakrishnan, which we describe below.

Although we can compare batsman by carefully considering their batting vectors, we choose to concentrate here on a few key players and leave the remainder to the reader. Table 5 gives the top 10 batsmen who scored a minimum of 300 runs based on the analysis of Madhusudhan Ramakrishnan. He is a sub-editor (stats) at ESPN cricinfo who has published "An analysis of individual batting and bowling performances in IPL 2012" (http://www.espncricinfo.com/indian-premier-league-2012/content/story/566523.html) subsequent to the IPL 2012 tournament. His method is based on a point system for each player, and he has provided a ranking of the top 10 batsman who scored a minimum of 300 runs along with a ranking of the top 10 bowlers who bowled more than 35 overs. We call this method the Ramakrishnan ranking method and compare it to the principal component method introduced here.

**Table 5.** Top Ten Batsmen for IPL 2012 (minimum 300 runs) using Ramakrishnan ranking method along with First PC Score $L_1$

| Batsman | Runs | Ave | SR | Ramakrishnan Score | First PC Score ($L_1$) |
|---|---|---|---|---|---|
| C.H. Gayle | 733 | 61.08 | 160.74 | 27.85 | 8.47 |
| G. Gambhir | 590 | 36.87 | 143.55 | 20.99 | 4.59 |
| K.P. Pietersen | 305 | 61.00 | 147.34 | 20.23 | 2.86 |
| C.L. White | 479 | 43.54 | 149.68 | 20.08 | 3.88 |
| S. Dhawan | 569 | 40.64 | 129.61 | 19.10 | 4.10 |
| V. Sehwag | 495 | 33.00 | 161.23 | 17.70 | 4.12 |
| F. du Plessis | 398 | 33.16 | 130.92 | 17.10 | 2.11 |
| A.M. Rahane | 560 | 40.00 | 129.33 | 16.93 | 4.00 |
| A.B. de Villiers | 319 | 39.87 | 161.11 | 16.58 | 2.31 |
| S.P.D. Smith | 362 | 40.22 | 135.58 | 14.88 | 1.23 |

It is interesting to note that nine of ten batsmen match in Tables 4 and 5, although the rankings vary slightly. For example, V. Sehwag is ranked higher than K. P. Pietersen in our principal component ranking, which is opposite of that in Ramakrishnan rankings. R.G. Sharma is ranked seventh in our list of top ten batsmen, but he does not appear in the Ramakrishnan top ten list. A deeper look shows that R. G. Sharma played 17 matches, scored 433 runs with batting average (Ave) 30.92, striking rate (SR) 126.6, one century, three half centuries, 39 boundaries (Fours) and 18 (Sixes). Sharma had only two "not out" cases, which is somewhat low compared to five "not out" cases for

A. B. de Villiers, who played 16 matches, scored 319 runs with average 39.87, striking rate 161.11, three half centuries , 26 boundaries and 15 sixes.

It is not surprising to see that C. H. Gayle is ranked number one by both methods. He scored 733 runs in 15 matches with average 61.08, striking rate 160.74, one century, 7 half centuries, and 59 sixes, and was undoubtedly the best batsman during the season. G. Gambhir also had a remarkable season and achieved the second highest ranking using each method. However, V. Sehwag has a third place ranking using the first principal component method, yet was ranked sixth using the Ramakrishnan ranking. Of course, V. Sehwag had a lower average of 33, while K.P. Pietersen had an average of 61. But, Pietersen has been "not out" three times and Sehwag has been "not out" only one time, which has inflated Pietersen's batting average. In terms of strike rate, Sehwag's 161.23 is much better than Pietersen's 147.34.

Another discrepancy between the two methods is evident in the case of Steven Smith. The first principal component ranking method does not place him in its top ten table. In fact, Table 6 shows his ranking to be twenty-three when considering all the batsmen and not just those who scored over 300 runs. Let's see why this is the case. Smith's striking rate and average are high, but he did not score even one half century. He has been "not out" five times, and that is why his average is that high. Because the first principal component method captures much information, it also penalizes for not having long innings, like centuries or half centuries. This demonstrates that the first principal component method perhaps takes into account more factors than might be considered under the Ramakrishnan method. In fact, since both methods select almost the same batsmen in their top 10 lists, the first principal component ranking technique is an interesting alternative to the ESPN ranking method of Ramakrishnan. Moreover, it is relatively easy to implement with standard software. For completeness, Table 6 gives the complete listing of first principal component scores and corresponding rankings for all ninety batmen.

**Table 6**. Complete List of IPL 2012 Batsmen Ranked by First Principal Component, $L_1$

| Batsman | $L_1$ | Rank | Batsman | $L_1$ | Rank | Batsman | $L_1$ | Rank |
|---|---|---|---|---|---|---|---|---|
| C.H. Gayle | 8.47 | 1 | D.R. Smith | 0.68 | 31 | Y.K. Pathan | -0.87 | 61 |
| G. Gambhir | 4.59 | 2 | K.A Pollard | 0.68 | 32 | S.T.R. Binny | -1.05 | 62 |
| V. Sehwag | 4.12 | 3 | M.E.K. Hussey | 0.67 | 33 | Harbhajan Singh | -1.17 | 63 |
| S. Dhawan | 4.10 | 4 | M.A. Agarwal | 0.53 | 34 | Y. V. Rao | -1.27 | 64 |
| A.M. Rahane | 4.00 | 5 | J.D. Ryder | 0.46 | 35 | R.E. Levi | -1.27 | 65 |
| C.L. White | 3.88 | 6 | M.S. Bisla | 0.32 | 36 | A.D. Mathews | -1.34 | 66 |
| R.G. Sharma | 2.90 | 7 | B.J. Hodge | 0.26 | 37 | N. Saini | -1.36 | 67 |
| K.P. Pietersen | 2.86 | 8 | N.V. Ojha | 0.25 | 38 | P.P. Chawla | -1.41 | 68 |
| A.B. de Villiers | 2.31 | 9 | B.B. McCullum | 0.09 | 39 | L.R. Shukla | -1.44 | 69 |
| F. du Plessis | 2.11 | 10 | A.C. Gilchrist | -0.05 | 40 | Shakib Al Hasan | -1.49 | 70 |
| D.A. Warner | 2.07 | 11 | D.B. Das | -0.20 | 41 | M.N. Samuels | -1.49 | 71 |
| J.P. Duminy | 2.07 | 12 | M.K. Tiwary | -0.29 | 42 | M.J. Clarke | -1.70 | 72 |
| O.A. Shah | 1.93 | 13 | Azhar Mahmood | -0.36 | 43 | R. Vinay Kumar | -1.91 | 73 |
| S.K. Raina | 1.86 | 14 | I.K. Pathan | -0.36 | 44 | R. Bhatia | -1.93 | 74 |
| R. Dravid | 1.82 | 15 | S. Badrinath | -0.38 | 45 | J. Botha | -1.97 | 75 |
| D.J. Hussey | 1.81 | 16 | J.E.C. Franklin | -0.54 | 46 | S.P. Goswami | -2.02 | 76 |
| Mandeep Singh | 1.79 | 17 | L.R.P.L. Taylor | -0.54 | 47 | P. Kumar | -2.05 | 77 |
| S.R. Watson | 1.77 | 18 | M.K. Pandey | -0.56 | 48 | A. Ashish Reddy | -2.05 | 78 |
| D.J. Bravo | 1.72 | 19 | S.C. Ganguly | -0.58 | 49 | S.L. Malinga | -2.08 | 79 |
| A.T. Rayudu | 1.44 | 20 | K.D. Karthik | -0.58 | 50 | D.L. Vettori | -2.11 | 80 |
| R.V. Uthappa | 1.38 | 21 | K.C. Sangakkara | -0.60 | 51 | R.J. Peterson | -2.20 | 81 |
| M. Vijay | 1.27 | 22 | R.A. Jadeja | -0.64 | 52 | B. Kumar | -2.21 | 82 |
| S.P.D. Smith | 1.23 | 23 | A.L. Menaria | -0.70 | 53 | R. Ashwin | -2.27 | 83 |
| S.E. Marsh | 1.17 | 24 | D.T. Christian | -0.72 | 54 | D.W. Steyn | -2.73 | 84 |
| J.H. Kallis | 1.09 | 25 | S.S. Tiwary | -0.75 | 55 | A. Mishra | -2.87 | 85 |
| D.P.M.D. Jayawardene | 1.04 | 26 | M. Manhas | -0.79 | 56 | W.D. Parnell | -2.98 | 86 |
| T.M. Dilshan | 1.03 | 27 | Y. Nagar | -0.79 | 57 | Z. Khan | -2.99 | 87 |
| V. Kohli | 1.02 | 28 | D.A. Miller | -0.80 | 58 | P.C. Valthaty | -3.05 | 88 |
| S.R. Tendulkar | 0.88 | 29 | P.A. Patel | -0.82 | 59 | R.P. Singh | -3.40 | 89 |
| M.S. Dhoni | 0.86 | 30 | J.A. Morkel | -0.83 | 60 | R. Sharma | -3.93 | 90 |

## 4.2 Ranking Bowlers using the First Principal Component

This analysis includes all bowlers who have bowled at least ten *overs*, which accounts for 83 total bowlers. The variables most useful for measuring the quality of bowlers are Wickets (**Wkts**), Bowling Average (**Ave**), Strike Rate (**SR**), and Economy Rate (**Econ**). For each bowler, values for each of these variables are collected together into a $(4 \times 1)^t$ column vector of the form (**Wkts**, **Ave**, **SR**, **Econ**)$^t$, which we call the *bowling vector*.

As with the batmen, we use the $(4 \times 4)$ sample correlation matrix associated with the sample bowling vectors in an attempt to understand the correlation structure of the bowling vectors. The correlation matrix is used to ensure standardization among the components of the bowling vector. As with batsmen, we use Minitab® for computing the eigenvalue-eigenvector pairs of the sample correlation matrix. Table 7 shows the sample correlation matrix for the eighty-three bowling vectors (for eighty-three bowlers).

**Table 7.** Sample Correlation Matrix for 83 Bowlers

| | **Wkts** | **Ave** | **Econ** | **SR** |
|---|---|---|---|---|
| **Wkts** | **1.00** | -0.49 | -0.29 | -0.51 |
| **Ave** | -0.49 | **1.00** | 0.52 | 0.96 |
| **Econ** | -0.29 | 0.52 | **1.00** | 0.33 |
| **SR** | -0.51 | 0.96 | 0.33 | **1.00** |

Table 8 provides the ordered eigenvalues and percentage of total variability attributed to each, while Table 9 gives the coefficients for all four principal component (PC1 – PC4). The eigenvalue-eigenvector pair for the first principal component is highlighted in Table 9.

**Table 8.** Ordered eigenvalues and corresponding Percentages of Total Variability for Bowlers

| Eigenvalue | **2.616** | **0.752** | **0.620** | **0.012** |
|---|---|---|---|---|
| Total Variability | 65.40% | 18.79% | 15.51% | 0.03% |

**Table 9.** Eigenvalue and Eigenvector Pairs for the Sample Correlation Matrix in Table 7.

| Eigenvalues: | **2.616** | 0.752 | 0.620 | 0.012 |
|---|---|---|---|---|
| Variable | **PC1** | PC2 | PC3 | PC4 |
| **Wkts** | **-0.428** | 0.335 | -0.839 | 0.038 |
| **Ave** | **0.591** | -0.048 | -0.354 | -0.723 |
| **Econ** | **0.383** | 0.892 | 0.168 | 0.172 |
| **SR** | **0.566** | -0.301 | -0.3789 | 0.668 |

Accordingly, the first principal component for bowlers is

$$L_1 = -0.428*\mathbf{Wkts} + 0.591*\mathbf{Ave} + 0.383*\mathbf{Econ} + 0.566*\mathbf{SR}$$

where the variables **Wkts**, **Ave**, **Econ**, **SR** have already been standardized. The first entry in Table 8 shows that 65.40% of the Total Variability can be explained by this first principal component alone. Its corresponding eigenvalue 2.616 is almost 3.5 times the magnitude of the next largest eigenvalue, 0.752. Again following Kaiser's rule (Kaiser 1960), we use the first principal component since its corresponding eigenvalue is the only one which is greater than one. The accompanying Scree plot is shown in Figure 8, which indicates an elbow at abscissa 2, implying that the first principal component is also sufficient for ranking bowlers.

**Figure 8**. Scree Plot for IPL 2012 Bowling.

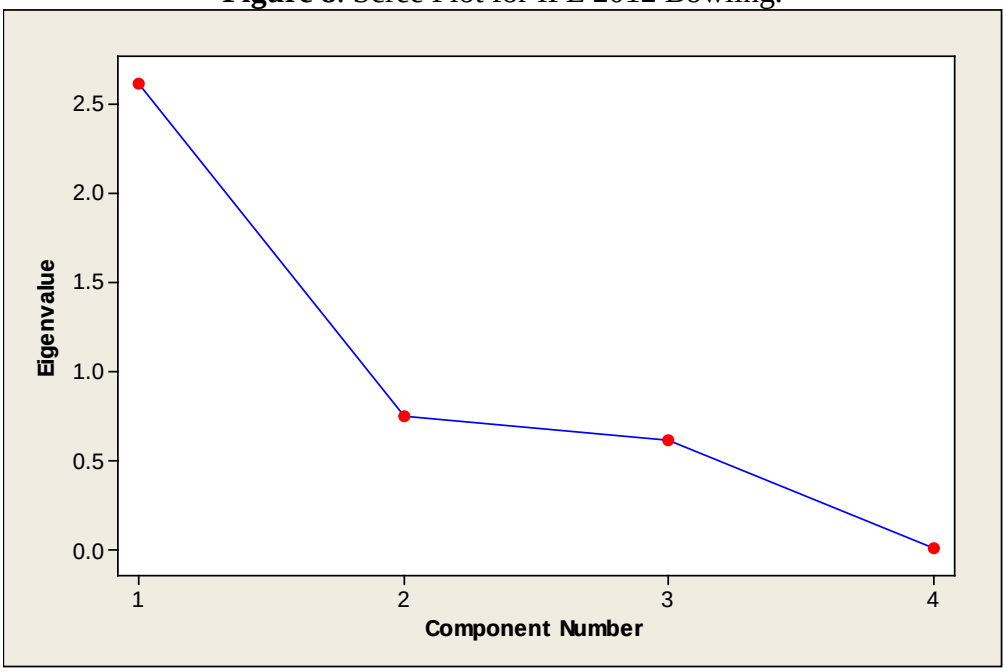



**Table 10.** Top Ten Bowlers for IPL-2012 (minimum 35 overs) using First Principal Component, $L_1$

| Bowler | Matches | Wickets | Ave | Econ | SR | $L_1$ |
|---|---|---|---|---|---|---|
| S.P. Narine | 15 | 24 | 13.50 | 5.47 | 14.70 | -2.92 |
| S.L. Malinga | 14 | 22 | 15.90 | 6.30 | 15.10 | -2.40 |
| M. Morkel | 16 | 25 | 18.12 | 7.19 | 15.10 | -2.27 |
| D.W. Steyn | 12 | 18 | 15.83 | 6.10 | 15.50 | -2.14 |
| M. Muralitharan | 10 | 15 | 17.33 | 6.50 | 16.00 | -1.71 |
| U.T. Yadav | 17 | 19 | 23.84 | 7.42 | 19.20 | -1.42 |
| P. Awana | 12 | 17 | 21.88 | 7.91 | 16.50 | -1.25 |
| K.A. Pollard | 14 | 16 | 21.87 | 7.98 | 16.40 | -1.15 |
| Z. Khan | 16 | 17 | 26.64 | 7.55 | 21.10 | -1.07 |
| P.P. Chawla | 16 | 16 | 26.18 | 7.35 | 21.30 | -1.07 |

Unlike the analysis for batsmen, where the coefficients in $L_1$ of that discussion are all positive, the coefficient of **Wkts** here is negative while the remainder are positive. This makes sense since better bowler performance is naturally associated with higher numbers of wickets taken from batsmen. On the other hand, lower values for the other three variables: Bowling Average (**Ave**), Economy Rate (**Econ**) and Bowling Strike rate (**SR**) indicate better bowler performance. Since $L_1$ = -0.428***Wkts** + 0.591***Ave** + 0.383***Econ** + 0.566***SR** is a type of weighted average of these four variables, it is reasonable is to use the first principle component for ranking. In this context, $L_1$ can reasonably be described as the *general-bowling-performance-index,* where smaller (negative) values indicate stronger bowler performance.

Table 10 lists the top ten bowlers using the first principal component method, while Table 11 gives the Ramakrishnan rankings. Both methods place the same players in their top five lists although the rank orderings are slightly different for some bowlers. When considering top ten bowlers using these methods, seven are common to both lists. Clearly, the simplicity of the first principal

component method makes it is very appealing in this context as well. For completeness, Table 12 presents the complete rankings of IPL 2012 bowlers with at least ten *overs.*

**Table 11.** Top 10 Bowlers for IPL-2012 (minimum 35 overs): Ramakrishnan Ranking Method along with First PC Score *L*

| Bowler | Matches | Wickets | Average | Economy rate | Ramakrishnan Score | First PC Score ($L_1$) |
|---|---|---|---|---|---|---|
| D.W. Steyn | 12 | 18 | 15.83 | 6.10 | 29.12 | -2.14 |
| S.P. Narine | 15 | 24 | 13.50 | 5.47 | 28.02 | -2.92 |
| M. Muralitharan | 10 | 15 | 17.33 | 6.50 | 27.67 | -1.71 |
| S.L. Malinga | 14 | 22 | 15.90 | 6.30 | 25.76 | -2.40 |
| M. Morkel | 16 | 25 | 18.12 | 7.19 | 24.75 | -2.27 |
| P. Awana | 12 | 17 | 21.88 | 7.91 | 23.60 | -1.25 |
| G.B. Hogg | 9 | 10 | 25.30 | 7.02 | 22.49 | -0.72 |
| Azhar Mahmood | 11 | 14 | 23.50 | 7.71 | 22.00 | -0.97 |
| Z. Khan | 16 | 17 | 26.64 | 7.55 | 20.86 | -1.07 |
| M.M. Patel | 12 | 15 | 24.46 | 7.86 | 20.80 | -0.96 |

**Table 12.** Complete List of IPL-2012 Bowlers (at least 10 overs) Ranked by First Principal Component, *L1*

| Bowler | Rank | $L_1$ | Bowler | $L_1$ | Rank |
|---|---|---|---|---|---|
| S.P. Narine | 1 | -2.92 | A.D. Mathews | -0.29 | 43 |
| S.L. Malinga | 2 | -2.40 | Pankaj Singh | -0.28 | 44 |
| M. Morkel | 3 | -2.27 | R.P. Singh | -0.19 | 45 |
| D.W. Steyn | 4 | -2.14 | J. Botha | -0.15 | 46 |
| L. Balaji | 5 | -1.84 | V. Pratap Singh | -0.08 | 47 |
| M. Muralitharan | 6 | -1.71 | K.P. Appanna | -0.02 | 48 |
| B.W. Hilfenhaus | 7 | -1.59 | Harmeet Singh | -0.01 | 49 |
| Shakib Al Hasan | 8 | -1.54 | R.E. van der Merwe | -0.01 | 50 |
| U.T. Yadav | 9 | -1.42 | D.T. Christian | 0.03 | 51 |
| A.B. McDonald | 10 | -1.36 | B. Kumar | 0.04 | 52 |
| P. Awana | 11 | -1.25 | R. Sharma | 0.05 | 53 |
| A.D. Mascarenhas | 12 | -1.22 | S.T.R. Binny | 0.05 | 54 |
| K.A. Pollard | 13 | -1.15 | A. Nehra | 0.12 | 55 |
| K. Cooper | 14 | -1.12 | A. Singh | 0.18 | 56 |
| Z Khan | 15 | -1.07 | M.J. Clarke | 0.23 | 57 |
| P.P. Chawla | 16 | -1.07 | S. Nadeem | 0.28 | 58 |
| R.J. Harris | 17 | -1.04 | S.B. Jakati | 0.28 | 59 |
| R. Vinay Kumar | 18 | -1.04 | Iqbal Abdulla | 0.29 | 60 |
| A. Chandila | 19 | -1.03 | V.R. Aaron | 0.31 | 61 |
| A.B. Dinda | 20 | -1.02 | H.V. Patel | 0.33 | 62 |
| Azhar Mahmood | 21 | -0.97 | A.A. Chavan | 0.35 | 63 |
| -M.M. Patel | 22 | -0.96 | A.C. Thomas | 0.61 | 64 |
| J.H. Kallis | 23 | -0.92 | Ankit Sharma | 0.65 | 65 |
| R.A. Jadeja | 24 | -0.83 | B. Lee | 0.66 | 66 |
| R. Ashwin | 25 | -0.81 | P. Kumar | 0.68 | 67 |
| P. Negi | 26 | -0.73 | D.L. Vettori | 0.90 | 68 |
| G.B. Hogg | 27 | -0.72 | Anand Rajan | 0.94 | 69 |
| S.K. Trivedi | 28 | -0.67 | D.R. Smith | 1.23 | 70 |
| R. Bhatia | 29 | -0.63 | M. de Lange | 1.30 | 71 |
| P.P. Ojha | 30 | -0.62 | A.B. Agarkar | 1.39 | 72 |
| D.J. Bravo | 31 | -0.58 | I.K. Pathan | 1.42 | 73 |
| A. Ashish Reddy | 32 | -0.57 | S.K. Raina | 1.55 | 74 |
| J.A. Morkel | 33 | -0.54 | S.C. Ganguly | 1.64 | 75 |

| | | | | | |
|---|---|---|---|---|---|
| A. Mishra | 34 | -0.52 | Harbhajan Singh | 1.86 | 76 |
| S.W. Tait | 35 | -0.51 | Y.K. Pathan | 2.02 | 77 |
| W.D. Parnell | 36 | -0.51 | M. Kartik | 2.17 | 78 |
| J.E.C. Franklin | 37 | -0.44 | B.A. Bhatt | 2.48 | 79 |
| D.E. Bollinger | 38 | -0.44 | M.S. Gony | 2.50 | 80 |
| K.M.D.N. Kulasekara | 39 | -0.43 | J.P. Duminy | 4.49 | 81 |
| M.N. Samuels | 40 | -0.35 | T.P. Sudhindra | 6.25 | 82 |
| S.R. Watson | 41 | -0.33 | A.D. Russell | 7.34 | 83 |
| P. Parameswaran | 42 | -0.32 | | | |

## 4. Conclusions

Quantifying athletic performance (performance analysis) is a challenging task in any sport. It is especially important in competitive sports impacted by player auctions or trades which, by their nature, usually involve organizations spending large monetary sums with the hope of future, competitive advantages. The basis of these transactions lies in past player performance, and there are typically several indicators or aspects available to measure the various contributions of prized athletes. Unfortunately, these indicators are generally highly correlated with one another, making it difficult to judge overall player performance. There is no doubt that currently available player ranking procedures are opaque, so there is a need for new and transparent methods of performance analysis.

Although expert opinion can be quite valuable; it is also very subjective. Here, we have demonstrated a simple method using principal component analysis that can be directly applied to correlated, multivariate data. Using 2012 Indian Premier League (IPL 2012) data, we have shown how to rank batsmen and bowlers based on their contributions to their teams during this competitive season. The simplicity and straight-forwardness of this technique make it very appealing as an introduction to the topic of principal component analysis. Moreover, this real classroom example is accessible to upper undergraduate and graduate students having just a basic understanding of multivariate statistics. As a result, the motivation and relevance of principal component analysis becomes immediately apparent to them, and real learning soon follows.

## References

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, 1266-1274.

Bennet, J. (1998), *Statistics in Sports*, New York, NY: Oxford University Press.

Borooah, V. K., and Mangan, J. E. (2010), "The 'Bradman Class': An Exploration of Some Issues in the Evaluation of Batsmen for Test Matches, 1877–2006.", *Journal of Quantitative Analysis in Sports*, 6(3), Article 14.

Dawkins, B. (1989), "Multivariate Analysis of National Track Records," *The American Statistician*, 43, 100-115.

Johnson, R. A., and Wichern, D. W. (2007*), Applied Multivariate Statistical Analysis* (6th ed.), Upper Saddle River, NJ: Prentice Hall.

Kaiser, H. (1960), "The Application of Electronic Computers to Factor Analysis," *Educational and Psychological Measurement*, 20, 141-151.

Lakkaraju, P., and Sethi, S. (2012), "Correlating the Analysis of Opinionated Texts Using SAS® Text Analytics with Application of Sabermetrics to Cricket Statistics," *Proceedings of SAS Global Forum 2012*, 136-2012, 1-10.

Lemmer, H. (2004), "A Measure for the Batting performance of Cricket Players," *South African Journal for Research in Sport, Physical Education and Recreation*, 26, 55-64.

Lemmer, H. (2008b), "An Analysis of Players' Performances in the First Cricket Twenty20 World Cup Series," *South African Journal for Research in Sport, Physical Education and Recreation* 30, 71-77.

Lemmer, H. (2012), "The Single Match Approach to Strike Rate Adjustments in Batting Performance Measures in Cricket," *Journal of Sports Science and Medicine*, 10, 630-634.

Lewis, A. (2008), "Extending the Range of Player-Performance Measures in One-Day Cricket," *Journal of Operational Research Society*, 59, 729-742.

Ramakrishnan, M. (2012), "Indian Premier League 2012 / Stats Analysis" [*ESPN* online], http://www.espncricinfo.com/indian-premier-league-2012/content/story/566523.html.

Van Staden, P. (2009), "Comparison of Cricketers' Bowling and Batting Performance using Graphical Displays," *Current Science*, 96**,** 764-766.

Watnik, M., and Levine, R. (2001), "NFL Y2K PCA," *Journal of Statistics Education* [online], 9, 3. Available at http://www.amstat.org/publications/jse/v9n3/datasets.watnik.html**.**

---

Ananda B. W. Manage
Department of Mathematics and Statistics
Sam Houston State University
420 Lee Drain Building
Huntsville, TX 77341-2206
Email: wxb001@shsu.edu

Stephen M. Scariano
Department of Mathematics and Statistics
Sam Houston State University
420 Lee Drain Building
Huntsville, TX 77341-2206
Email: scariano@shsu.edu

---

Volume 21 (2013) | Archive | Index | Data Archive | Resources | Editorial Board | Guidelines for Authors | Guidelines for Data Contributors | Guidelines for Readers/Data Users | Home Page | Contact JSE | ASA Publications