Chapter 2
Principal Component Analysis in Transfer Function

Abstract This chapter explores the transfer function in detail, with multiple linear regressions, and principal component analysis (PCA). Furthermore, it contains the slight description of various types of regression and emphasizes on the PCA and the calculations of principal components (PCs) in detail.

Keywords Transfer function · Regression methods · PCA · PCR

2.1 Transfer Function/Regression Method

The most popular approach of downscaling is the use of transfer function which is a regression-based downscaling method. The transfer function method relies on direct quantitative relationship between the local scale climate variable (predictand) and the variables containing the large scale climate information (predictors) through some form of regression. Individual downscaling schemes differ according to the choice of mathematical transfer function, predictor variables, or statistical fitting procedure. To date, linear and nonlinear regression, artificial neural network (ANN), canonical correlation, etc., have been used to derive predictor–predictand relationship. Among them, ANN-based downscaling techniques have gained wide recognition owing to their ability to capture nonlinear relationships between predictors and predictand. The main strength of transfer function downscaling is the relative ease of application. The main weakness is that the models often explain only a fraction of the observed climate variability (especially in precipitation series). Transfer methods also assume validity of the model parameters under future climate conditions. The downscaling is highly sensitive to the choice of predictor variables and statistical form. The schematic diagram of transfer function is given in Fig. 2.1.

Large-scale values of particular climate variables (predictors) will be used to predict the values of the site-specific variables (predictand). The large-scale area should roughly correspond to the size of the GCM grid box. It may be necessary to

T.M.V. Suryanarayana and P.B. Mistry, *Principal Component Regression for Crop Yield Estimation*, SpringerBriefs in Applied Sciences and Technology,
DOI 10.1007/978-981-10-0663-0_2

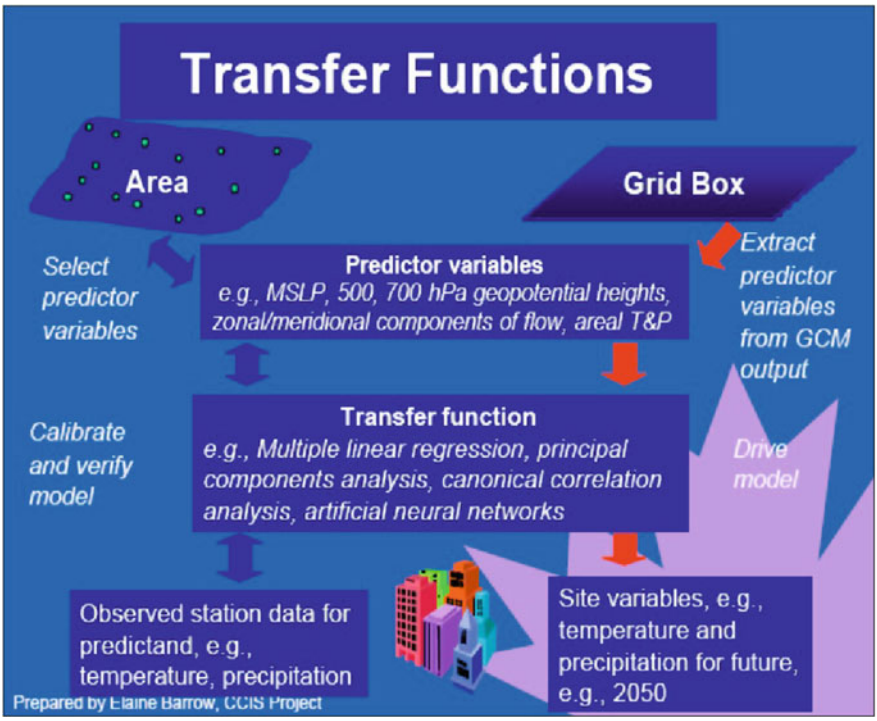

Fig. 2.1 Schematic diagram of transfer function

construct area-average values of, say, mean temperature or precipitation (usually simple averaging of station data, or weighted averaging). First step is to define the predictor variables—they must explain a high proportion of the variance in the predictand and then construct the transfer function relating the site-specific variable to the larger-scale predictors using an appropriate technique—being aware of the constraints associated with the method being used. For example, in multiple linear regressions it is assumed that the predictor variables are independent, i.e., the correlation between them is effectively zero. If this is not the case then the regression coefficients will not be a true estimate of the contribution of each of the predictor variables to the variance of the predictand. Keep back some data in order to test the performance of the model (validation).To derive the predictand values under a future climate, the larger-scale predictors derived from GCM data are used to drive the transfer function. The main advantages are firstly it is computationally much less demanding and secondly, ensembles of high resolution climate scenarios may be produced relatively easily (Barrow 2001).

2.2 Types of Regressions

2.2.1 The Simple Linear Regression Model

The relationship between a response variable Y and a predictor variable X is postulated as a linear model

$$Y = b_0 + b_1 X + E$$

where b_0 and b_1, are called the model regression coefficients, and E is a random disturbance or error. It is assumed that in the range of the observations studied, the linear equation above provides an acceptable approximation to the true relation between Y and X. In other words, Y is approximately a linear function of X, and E measures the discrepancy in that approximation.

In particular, E contains no systematic information for determining Y that is not already captured in X. The coefficient b_1, called the slope, may be interpreted as the change in Y for unit change in X. The coefficient b_0, called the constant coefficient or intercept, is the predicted value of Y, when $X = 0$.

2.2.2 The Multiple Linear Regression Model

Multiple linear regressions are with two or more independent variables on the right-hand side of the equation. Multiple linear regressions should be used, if more than one cause is associated with the effect, one wish to understand.

The equation and the true plane for the case of two independent variables, we can write the equation for a multiple regression model this way:

$$Y = \alpha + \beta X + \gamma Z + \text{error}$$

Imagine that the X- and Z-axes are on a table in front of you, with the X-axis pointing to the right and the Z-axis pointing directly away from you. The Y-axis is standing vertically, straight up from the table.

$Y = \alpha + \beta X + \gamma Z + \text{error}$ is the formula for a flat plane, that is floating in the three-dimensional space.

α is the height of the plane above the point on the table where $X = 0$ and $Z = 0$
β is the slope of the plane in the X direction, how fast the plane rises as one go to the right

If you have more than two independent variables, it is conventional to go to a subscript notation for the variables and the slope parameters as given in the equation given below.

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 \ldots + \text{error}$$

2.2.3 Polynomial Regression Models

A model is said to be linear when it is linear in parameters. So the model

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \text{error}$$

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{12} X_1 X_2 + \text{error}$$

are also the linear model. In fact, they are the second order polynomials with one and two variables, respectively.

The polynomial models can be used in those situations, where the relationship between the study and the explanatory variables is curvilinear. Sometimes, a nonlinear relationship in a small range of explanatory variables can also be modeled by the polynomials.

The *K*th order polynomial model with one variable is given by the below given equation.

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_k X^k + \text{error}$$

2.2.4 Nonlinear Regression

In statistics, nonlinear regression is a form of analysis, in which observational data are modeled by a function, which is a nonlinear combination of the model parameters and depends on one or more independent variables. The data are fitted by a method of successive approximations.

2.3 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) involves a mathematical procedure that transforms a number of possibly correlated variables into a smaller number of uncorrelated variables called PC. Principal components are also known by generation of a new set of variables by PCA, and they have a linear combination of the

original variables. In this study, due to large dimensionality of predictor variables, it may be computationally unstable. Hence, PCA is performed to reduce the dimensionality of the predictor variables. PCA is also used to downscale GCM outputs of large-scale climatic variables to sub-divisional level.

> PCA is a way of identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences. Since patterns in data can be hard to find in data of high dimension, where the luxury of graphical representation is not available, PCA is a powerful tool for analyzing data.

Often, the variables under study are highly correlated and as such they are effectively "saying the same thing". It may be useful to transform the original set of variables to a new set of uncorrelated variables called PC. These new variables are linear combinations of original variables and are derived in decreasing order of importance so that the first principal component accounts for as much as possible of the variation in the original data. Also, PCA is a linear dimensionality reduction technique, which identifies orthogonal directions of maximum variance in the original data, and projects the data into a lower-dimensionality space formed of a subset of the highest variance components (Agrawal and Rao).

2.3.1 Advantages and Disadvantages of PCA

Principal component analysis (PCA) is a standard tool in modern data analysis in diverse fields from neuroscience to computer graphics—because it is a simple, nonparametric method for extracting relevant information from confusing data sets. With minimum effort, PCA provides a roadmap for how to reduce a complex data set to a lower dimension to reveal the sometimes hidden, simplified structures that often underlie it.

Another field of use is pattern recognition and image compression, therefore PCA is suited for use in facial recognition software for example, as well as for recognition and storing of other biometric data. PCA is also used in research of agriculture, biology, chemistry, climatology, demography, ecology, food research, genetics, geology, meteorology, oceanography, psychology, quality control, etc. PCA has been used in economics and finance to study changes in stock markets, commodity markets, economic growth, exchange rates, etc. Earlier studies were done in economics, but stock markets were also under research already. "Principal component or factor analysis has been used in several recent empirical studies concerned with the existence of general movements in the returns from common stocks."

PCA is a special case of factor analysis that is highly useful in the analysis of many time series and the search for patterns of movement common to several series (true factor analysis makes different assumptions about the underlying structure and

solves eigenvectors of a slightly different matrix). This approach is superior to many of the bivariate statistical techniques used earlier, in that it explores the interrelationships among a set of variables caused by common "factors," mostly economic in nature. PCA is a way of identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences. A primary benefit of PCA arises from quantifying the importance of each dimension for describing the variability of a data set. PCA can also be used to compress the data, by reducing the number of dimensions, without much loss of information. When using PCA to analyze a data set, it is usually possible to explain a large percentage of the total variance with only a few components. Principal components are selected so that each successive one explains a maximum of the remaining variance; the first component is selected to explain the maximum proportion of the total variance, the second to explain the maximum of the remaining variance, etc. PCA is completely nonparametric: any data set can be plugged in and an answer comes out, requiring no parameters to tweak and no regard for how the data was recorded. From one perspective, the fact that PCA is nonparametric (or plug-and-play) can be considered a positive feature because the answer is unique and independent of the user (Kumar and chauhan 2014).

2.3.2 Applications of Principal Components

The most important use of PCA is reduction of data. It provides the effective dimensionality of the data. If first few components account for most of the variation in the original data, then first few components' scores can be utilized in subsequent analysis in place of original variables.

Plotting of data becomes difficult with more than three variables. Through PCA, it is often possible to account for most of the variability in the data by first two components, and it is possible to plot the values of first two components scores for each individual. Thus, PCA enables us to plot the data in two dimensions. Particularly, detection of outliers or clustering of individuals will be easier through this technique. Often, use of PCA reveals grouping of variables, which would not be found by other means.

Reduction in dimensionality can also help in analysis, where number of variables is more than the number of observations, for example, in discriminant analysis and regression analysis. In such cases, PCA is helpful by reducing the dimensionality of data.

Multiple regressions can be dangerous, if independent variables are highly correlated. PCA is the most practical technique to solve the problem. Regression analysis can be carried out using PC as regressors in place of original variables. This is known as principal component regression (Agrawal and Rao).

2.4 Principal Component Regression (PCR)

In order to conduct principal component regression, PCA is carried out as follows:

Principal Component Analysis
When starting a research, students as well as researchers, often collect a lot of data or sometimes come across large datasets that are available. But when having lots of data, especially when it is secondary data, it is often very easy to get confused. It is hard to find the variables that are really important for the research when there are so many variables to consider. This is where principal components analysis (PCA) can help.

Principal components analysis (PCA) was invented by Karl Pearson in 1901 and is now used in many fields of science. PCA is mostly used as a tool in exploratory data analysis because what it essentially does is to find the most important variables (a combination of them) that explain most of the variance in the data. So, when there is lots of data to be analyzed, PCA can make the task a lot easier. PCA also helps to construct predictive models (Chen et al. 2011).

2.4.1 Calculating Principal Components

The Principal components (PCs) can be found using purely mathematical arguments—they are given by an orthogonal linear transformation of a set of variables optimizing a certain algebraic criterion. An overview how to perform PCA is given hereafter.

Let $x_1, x_2, x_3 \ldots x_p$ are variables under study, and then first principal component may be defined as

$$z_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1P}x_P$$

Such that variance of z_1 is as large as possible subject to the condition that

$$a_{11}^2 + a_{12}^2 + \cdots a_{1p}^2 = 1$$

This constraint is introduced because if this is not done, then Var (z_1) can be increased simply by multiplying any $a_{1\text{js}}$ by a constant factor. The second principal component is defined as

$$z_2 = a_{21}x_1 + a_{22}x_2 + \cdots + a_{2P}x_P$$

Such that Var (z_2) is as large as possible next to Var (z_1) subject to the constraint that

$$a_{21}^2 + a_{22}^2 + \cdots + a_{2p}^2 = 1 \text{ and } \quad \text{cov}(z_1, z_2) = 0 \text{ and } \quad \text{so on.}$$

It is quite likely that first few PC accounts for most of the variability in the original data. If so, these few PC can then replace the initial p variables in subsequent analysis, thus reducing the effective dimensionality of the problem. An analysis of PC often reveals relationships that were not previously suspected and thereby allows interpretation that would not ordinarily result. However, PCA is more of a means to an end rather than an end in itself because this frequently serves as intermediate steps in much larger investigations by reducing the dimensionality of the problem and providing easier interpretation. It is a mathematical technique, which does not require user to specify the statistical model or assumption about distribution of original variants. It may also be mentioned that, PCs are artificial variables and often, it is not possible to assign physical meaning to them. Further, since PCA transforms original set of variables to the new set of uncorrelated variables, it is worth stressing that, if original variables are uncorrelated, and then there is no point in carrying out PCA (Agrawal and Rao).

2.4.2 Rules for Retaining Principal Components

In the initial extraction process, PCA will derive as many components as the number of measured variables. After the initial components are extracted, the analyst must decide on how many components should be retained to meaningfully represent the original correlation matrix. The initial component eigenvalues, percent of variance accounted for, and cumulative variance accounted for are provided. According to Stevens, "Probably the most widely used criterion is that of: Retain only those components whose eigenvalues are greater than 1".

A fairly common technique noted in the literature combines the two approaches. Eigen values greater than one is initially retained and the screen test is used subsequently to assess the tenability of the model. Because Eigen values represent reproduced variance, this is equivalent to setting a minimum level of acceptable variance reproduced by a component. The second stage evaluates the parsimony of the solution relative to the contribution of each component to reproducing the original variance in the data. A potential disadvantage of this approach is the arbitrary criterion of retaining eigenvalues greater than one in the first stage. Because PCA studies typically rely on sample data, eigenvalues (reproduced variance) should be expected to change (even with large samples) slightly from sample to sample. In addition, the interpretation of what constitutes a "meaningful" amount of variance accounted for (which eigenvalues represent) is inherently subjective (Kellow 2006).

2.4.3 Development of Principal Component Regression (PCR)

Principal component regression (PCR) is a type of regression analysis, which considers PC as independent variables, instead of adopting original variables. The PCs are the linear combination of the original variables which can be obtained by PCA. The PCA transforms the original set of intercorrelated independent variables to a new set of uncorrelated variables (i.e., PCs). The use of these PCs as independent variables is quite useful in the multiple regression models to avoid the multicollinearity problem and to identify the variables which are the most significant in making the prediction. The PCR models have been developed using PCs as inputs to predict and to compare the same with multiple linear regression models. It has been found that that the incorporation of PCs as independent variables in the regression models improved the model prediction as well as reduced the model complexity by eliminating multicollinearity.

Principal components regression (PCR) is a method for combating multicollinearity and results in estimation and prediction better than ordinary least squares, when used successfully. With this method, the original k climatic variables are transformed into a new set of orthogonal or uncorrelated variables called PC of the correlation matrix. This transformation ranks the new orthogonal variables in order of their importance and the procedure, and then involves eliminating some of the PC to effect a reduction in variance (Fekedulegn et al. 2002).

http://www.springer.com/978-981-10-0662-3

Principal Component Regression for Crop Yield Estimation

Suryanarayana, T.M.V.; Mistry, P.B.

2016, XVII, 67 p. 12 illus. in color., Softcover
ISBN: 978-981-10-0662-3