

CHAPTER 7

STATISTICAL MODELLING

SIMON J. MASON

International Research Institute for Climate and Society

OMAR BADDOUR

Moroccan Meteorological Service

Statistical models provide an alternative approach to using dynamical models in seasonal climate forecasting. In statistical models relationships between one set of data, the predictors, and a second set, the predictands, are sought. Common predictands include seasonal mean temperatures and accumulated precipitation, and are typically predicted using antecedent sea surface temperatures primarily within the tropical oceans. Predictions are made on the assumption that historically observed relationships are expected to apply in the future. There are many conditions for such an assumption to be valid, including the need for high-quality datasets to ensure that the historical relationships are robustly measured, and the need for relationships to have a sound theoretical basis. Because of the possibility of identifying spurious relationships between the predictors and the predictands, the statistical model should be tested carefully on independent data. Most statistical models are based on linear regression, which provides a “best guess” forecast under the assumption that a given change in the value of a predictor results in a constant change in the expected value of the predictand regardless of the value of the predictor. Modifications to the linear model can be made or alternative statistical procedures used when there is good reason to expect a relationship to be non-linear. However, other weaknesses of linear regression may also require these alternatives to be considered seriously. The primary problems with linear regression are multiplicity, multicollinearity, and non-normality of the predictands. Multiplicity refers to the effects of having a large number of candidate predictors: the danger of finding a spurious relationship increases. Multicollinearity arises when more than one predictor is used in the model and there are strong relationships between the predictors which can result in large errors in calculating the parameters of the model. Finally, a linear regression model may not be adequately constructed if the data being predicted have a strongly skewed or otherwise non-Gaussian distribution; seasonally accumulated

precipitation often exhibits such problems. Alternative forms of linear and non-linear statistical models can be applied to address such distributional problems.

7.1 Introduction

Whereas seasonal climate prediction using general circulation models is based upon successful modelling of the physics of the interactions between the atmosphere and the earth's surface (primarily the sea surface) and of the dynamics of these components of the climate system (Chapters 3 to 6), the earliest scientific efforts at forecasting seasonal climate anomalies were based on empirical observations of the atmosphere alone. In the late-nineteenth and early-twentieth centuries, Gilbert Walker, working on the problem of predicting the Indian monsoon, discovered that seasonal anomalies in different parts of the tropics were connected. For example, droughts in India and Australia would often occur in the same year. In such cases where there is a lag between the observed climate of one region and that of another, prediction may be possible. The most important pattern of connected climate anomalies identified by Walker was the Southern Oscillation, which describes opposite changes in sea-level pressure between the western and eastern Pacific Ocean, and involves major disruptions to the trade winds across the southern Pacific. Such relationships between climate anomalies in different areas are known as "teleconnections", and constituted the basis for early empirical methods of seasonal climate forecasting.

Teleconnections are suggestive of some large-scale forcing of the atmosphere, but it has only been since about the mid-1960s that forcing mechanisms have been identified and understood. The Southern Oscillation, for example, is closely related to the state of the sea surface temperatures (SSTs) in the equatorial Pacific Ocean: occasional large-scale warming and cooling of the equatorial Pacific Ocean, known as El Niño and La Niña respectively, simultaneously require and cause prolonged changes in the trade winds over the Pacific Ocean. These changes are associated with large-scale shifts in the location of areas of heavy rainfall, and, in turn, can affect climate conditions in other parts of the globe. Anomalous SSTs outside of the equatorial Pacific also can affect regional climate (for example, changes in the meridional SST gradient of the tropical Atlantic Ocean have important implications for rainfall over north-eastern Brazil and over much of West Africa). Most of the statistical prediction models used currently in operational forecasting attempt to model such relationships between observed climate and anomalous SSTs.

In this chapter, the basic principles of statistical modelling for seasonal climate prediction are introduced in section 7.2. Section 7.3 discusses in some detail the mathematics of linear regression, which is the most commonly used statistical prediction method used in practice. Linear regression forms the basic framework for a range of more sophisticated statistical techniques, and these, and other statistical techniques, are introduced in section 7.4, after a discussion of some of the limitations of linear regression.

7.2 Statistical modelling for climate prediction

Although some statistical seasonal climate prediction systems are built upon observed atmospheric teleconnections, the most common approach is to model historical relationships between the climate anomalies to be predicted and the underlying forcing mechanisms – specifically, observed SST anomalies. Statistical methods have been used by centres such as the Met Office (United Kingdom), the Bureau of Meteorology (Australia), and the National Centers for Environmental Prediction (USA) for a number of decades, and supplement the dynamically-based models that these centres also use. In the late 1990s, facilitated by extensive capacity building programs and an increasing availability of computing power, statistical methods of seasonal forecasting have been adopted by many national meteorological services throughout the world. These statistical models are constructed primarily to generate forecasts of seasonal precipitation totals, but air temperature forecasts are made also.

7.2.1 REQUIREMENTS FOR APPLYING STATISTICAL METHODS IN CLIMATE PREDICTION

Statistical methods aim to identify relationships between two sets of variables through statistical analyses performed on the historical records of the data known as time series. The two sets of variables are:

- a set of variables to be predicted (often denoted Y), and called predictands or response/dependent variables, such as seasonal total rainfall, and monthly average maximum and minimum temperatures;
- a set of variables used to make the predictions (often denoted X), and called predictors or explanatory/independent variables, such as SSTs or atmospheric indices (e.g. Southern Oscillation Index - SOI).

The intention is to identify within the historical records a “significantly” consistent relationship between observed values of the predictors and of the predictands. A “significantly” consistent relationship is one that is strong enough to be unlikely to have occurred by chance, and so provides a rea-

sonable level of confidence with which to make a prediction. Of course, for a prediction to be made, a lag between the observations on the predictors and on the predictands is implicit. The lag defines the lead-time of the forecast: by convention, the lead-time is defined as the time period between the end of the recording time of the predictors and the beginning of the target period. For example, if the average SSTs for June are used to predict the total rainfall for the three-month period August–October, the lag is 1-month (the last observation of the SSTs is made on 30 June, and the target period starts on 01 August). For any significant (lagged) relationship between the predictors and the predictands to be identified, there are some basic data requirements that must be met. These requirements are described in the following sections.

7.2.1.a Data quality issues

If relationships between predictors and predictands are to be modelled reliably, both sets of data need to be of high quality. The quality of a dataset is determined by the accuracy of the recorded values, the spatial and temporal resolution of the data, and the length of available records.

Apart from the problems of human and instrumental errors in recording climate variables, inaccuracies in historical records can arise from changes in instrumentation, relocation of recording sites, and/or changes in the recording environment. For example, the relocation of a thermometer even just a short way down slope could introduce an artificial jump in recorded temperatures because of adiabatic effects and changes in exposure. Any such changes in the recorded climate that are not a reflection of real changes are known as “inhomogeneities”. Statistical models are designed to “explain” the observed variability in the predictand data by reference to the observed variability in the predictor data. If part of the variability in the predictand dataset is a result of inhomogeneities, the statistical model will try to “explain” this component as if it were real. Similarly, if part of the variability in the predictor dataset is a result of inhomogeneities, the statistical model will try to use this component of the variability to “explain” the variability in the predictands. Correction for inhomogeneities is therefore an important component of the statistical model-building procedure. There are a variety of checks for data inhomogeneities, the most reliable of which make use of metadata. Metadata are information about the data themselves, and include, for example, information about any changes in instrumentation or changes in the location of the recording site.

Inhomogeneities in data can also be introduced by changes in the temporal resolution of the recordings. For example, the introduction of continuous temperature recordings has allowed a more accurate calculation of the daily mean temperature than was previously possible using only the

average of the maximum and minimum temperatures. The average of the maximum and the minimum tends to be higher than the integrated average, and so a change in the way the daily average is calculated could introduce an artificial change in the computed temperature. The temporal resolution of the data can also affect the quality of the information that can be communicated as part of a seasonal climate forecast. For example, although seasonal precipitation forecasts are usually communicated as some form of information about the total rainfall to be expected over a three-month period, if higher resolution data are available it may be possible to provide some information about the statistics of weather within the season. There are strong relationships between seasonal rainfall totals and rain-day frequencies and heavy rain-day frequencies in many parts of the world, and so a forecast of above-normal seasonal rainfall could be translated into statements about the numbers of days of rain (or heavy rain) that might be expected. However, these additional details are possible only if precipitation measurements are available at the daily timescale.

In addition to the temporal resolution, the spatial resolution of the data is of direct relevance to data quality issues. Station-based data, for example, are site specific, and forecasts that have been derived from models using station data may not be applicable to neighbouring areas. For precipitation, the applicability of a forecast for a nearby site can decline much more rapidly over short distances compared to that for temperature because of the highly localised nature of precipitation, especially in areas of convective rainfall. For precipitation forecasts, therefore, a relatively high density of stations would be advantageous. Sometimes forecasts are made for area-averaged precipitation or temperature. The area-averaging generally improves the forecast performance because the locally specific and unpredictable component of variability is reduced by the averaging. A downside, however, is that the forecast loses its specificity for individual locations, and so some form of translation is required to make the forecast relevant for specific locations. This translation is known as “downscaling” (see [Chapter 8](#)).

Other aspects of data quality, such as the presence of missing values and outliers, relate directly to sampling issues, and are discussed separately in the following section.

7.2.1.b Sampling issues

The extent to which a modelled relationship between predictors and predictands accurately represents the true relationship depends in part upon the number of records available. Inevitably there will be some errors in estimating the form and strength of this relationship because of the limited number of years for which climate observations are available, and such errors will

contribute to inaccurate predictions. These errors typically are larger for short records than for long records. For most statistical models used in seasonal climate forecasting it is recommended that at least 30 years of data be available for constructing a model in order to reduce the effects of sampling errors to an acceptable level.

There are three kinds of sampling errors that can occur when constructing a statistical model:

- the wrong predictors are selected;
- the wrong forms of the individual relationships between each predictor and the predictands are selected;
- the strength of the individual relationships between each predictor and the predictands is estimated incorrectly.

In practice, as the complexity of the model is increased each of the three forms of sampling error become more severe, and sample sizes need to be increased to compensate. To guard against the first two forms of error, statistical significance tests are performed as an attempt to estimate the probability that the error in question has occurred (i.e. that a spurious relationship has been identified). Because these tests are not foolproof, and are subject to problems (section 7.4.1), they should always be supplemented by theoretical considerations; a sound physical explanation should accompany any relationship that is implied by a statistical model. The theoretical basis can be supplied by research using GCMs, and/or by more detailed statistical analyses, perhaps using other climate datasets to investigate moisture fluxes, for example.

The poor availability of sufficient historical data to construct a robust statistical model is compounded by the presence of missing values. The simplest option is to omit the cases in which there are missing values from the analysis, but this approach easily can leave few or no cases with which to construct a model. Instead, attempts can be made to estimate the missing values. These procedures typically rely on relationships between various climatological variables. For example, if SSTs are to be used as predictors missing SST records could be estimated either from records for nearby locations and the spatial correlation structure of the temperatures, and/or from records immediately prior to and subsequent to the missing values and the temporal correlation structure for that location. Alternatively, if rainfall data are to be used as predictands, missing rainfall values could be estimated from the observed values for neighbouring stations, and/or from station values for variables that are not missing, such as temperature and humidity.

An additional aspect of sampling problems that should be addressed is the presence of outliers. Outliers are values either that are extreme in their own right, lying well outside of the range of the majority of the other data

records, or are values that are inconsistent with relationships with other variables. In either case, it has to be decided whether the outliers accurately represent what really happened because if they are retained they will have a large effect on most statistical models. If the outliers are considered accurate, it may still be desirable to reduce their impact on the model so that the data assumptions implicit in constructing the model are not violated (see further discussion in sections 7.3.3 and 7.4.1). For example, seasonal precipitation data for many parts of the globe are positively skewed⁴⁵; the largest seasonal totals therefore can have an undue influence on many statistical models, and this influence can be reduced by applying the model to the logarithms of the precipitation totals. The logarithmic transformation is often effective in reducing the positive skewness of data.

7.2.1.c Trends

Before attempting to build a statistical prediction model, it is common practice to remove any long-term trends in both the predictors and predictands. The argument for removing the trends is that if trends are present in the predictand(s) and any of the predictors the probability of identifying a spurious empirical relationship is increased. Effectively, the assumption of independent model errors is violated (section 7.3.3) unless the trends are removed. However, there are two situations under which it would be unadvisable to remove the trends: if there are prior reasons for expecting trends in the predictands to be caused by trends in any of the predictors; if trends are present in any of the predictands or of the predictors, but not in both. In the latter case, if there is a trend in a predictor, but not the predictand, it seems unreasonable to expect the higher frequency variability of the predictor to provide predictive skill, but for the long-term trend to be unrelated to the predictand; if there is a trend in the predictand, then a good statistical model would seek a predictor for this trend.

7.3 Building a statistical prediction model

In this section the primary steps in constructing a statistical model for climate prediction are detailed. The focus is on using SSTs as predictors and seasonal rainfall totals as predictands, although the procedure is similar for other variables. Linear regression modelling is used as a statistical model, while alternative statistical procedures are considered in section 7.4.2.

⁴⁵ Positive skewness occurs fairly commonly in meteorological data, and is evident in seasonal precipitation totals for many parts of the globe, most notably in arid and semi-arid areas. Maximum air temperatures in continental interiors can be weakly negatively skewed.

7.3.1 DEFINITION OF PREDICTANDS

Assuming that the necessary data quality control has been conducted, the first step in constructing a statistical model for seasonal climate prediction is to define the predictand. Seasonal rainfall totals are by far the most commonly used predictand, although increasing attention is being given to prediction of the intra-seasonal statistics of seasonal rainfall, such as the number of rain-days. Only one seasonal total per year is used in the model; other seasons are modelled separately because of the seasonally varying nature and influence of the forcing mechanisms that make seasonal climate prediction possible. The standard procedure is to define a season as a three-month total or average, but care should be taken to ensure that the season is defined appropriately; specifically, within a season the predictand should have a consistent response to the underlying forcing mechanisms. For example, in much of southern Africa, rainfall in November is positively associated with warm ENSO events, but the relationship in December and January is negative. It would therefore be inappropriate to forecast a November–January season.

If forecasts are to be made for regional averages rather than individual stations, the regions need to be delimited. The regions should be defined on the basis of similar relationships with the forcing mechanisms (for example, similar correlations with SSTs). There are numerous ways of defining the regions, and no single method has been identified as universally preferable. The most commonly used techniques include grouping stations with highest loadings on the same principal component (see section 7.4.2 for further discussion of principal components), and cluster analysis. Once stations have been allocated to a region, a regional rainfall index, r_k^* , is then calculated for each year, k , typically using the following equation:

$$r_k^* = \sum_{i=1}^m w_i \frac{r_{k,i} - \bar{r}_i}{s_i}, \quad (7.1)$$

where w_i is a weight applied to the i^{th} of m stations, $r_{k,i}$ is the rainfall at this i^{th} station during year k , and \bar{r}_i and s_i are the average and standard deviation of the station's rainfall, preferably calculated over a common reference period. The weights are defined to sum to unity, and can be set to avoid favouring unduly the contributions of clusters of stations to the regional index. In practice, if the station network is reasonably even, for the sake of simplicity the weights often are set equal for each station. The subtraction of the mean and division by the standard deviation standardises the data at each station and is designed to avoid giving stations with large mean and variance excessive weight (See [Chapter 8, section 8.3.3](#), for further discussion about standardisation, including some of its limitations).

7.3.2 DEFINITION OF CANDIDATE PREDICTORS

The most commonly used predictors in statistical models for seasonal climate prediction are SSTs. There are a number of global SST datasets available with varying spatial resolution (from $10^\circ \times 10^\circ$ to $1^\circ \times 1^\circ$), and some extend as far back as the mid-nineteenth Century (although data quality is considerably improved from about the 1950s). Whichever dataset is used, there are a large number of grids from which to choose, and some kind of pre-selection of grids and area-averaging of SSTs should be performed. Some area-averages have been predefined, such as the Niño3 index (5°S to 5°N , 150° to 90°W), but similar averages may be required for other areas if SSTs here are thought to have an important effect on rainfall variability in the region of interest. These area-averages should be defined based on theoretical considerations and extensive supporting statistical research. Simple correlations between the rainfall index and global SSTs followed by delimitation of areas with high correlation should be avoided because of problems with fishing (section 7.4.1) and subsequent problems of potential overestimation of the performance of the statistical model.

The temporal resolution of the predictors is not necessarily the same as that of the predictands. Because SSTs change much more slowly than the atmosphere, a one-month average is less noisy than a one-month average of some atmospheric variable, and more faithfully highlights recent trends in temperatures compared to a three-month average. As a result, statistical models are frequently constructed using SSTs for the latest month available. Of course, for an operational forecast to be made, the predictor data must be available before the beginning of the target period. The lag between the availability of the predictor data and the beginning of the target period defines the lead-time of the forecast (section 7.2.1).

7.3.3 STATISTICAL MODEL CONSTRUCTION

7.3.3.a Model formulation – simple linear regression

The simplest statistical model consists of a single predictand and a single predictor. In this case a regression model assumes a linear relationship between the predictor, x , and the predictand, y :

$$y = \beta_0 + \beta_1 x + \varepsilon, \quad (7.2)$$

where β_0 and β_1 are parameters to be estimated, and ε is an “error” term representing the unpredictable component of the predictand. The parameter β_0 is often called the “regression constant” or the “intercept”, while β_1 is referred to as the “regression coefficient” or the “slope”. The predictable component, \hat{y} , is given by:

$$\hat{y} = \beta_0 + \beta_1 x. \quad (7.3)$$

The objective in fitting a regression model is to estimate the parameters β_0 and β_1 so that the differences, or “residuals”, between the estimated⁴⁶ values of the predictands, \hat{y} , and the observed values, y , are minimised. From Eqs. (7.2) and (7.3):

$$\begin{aligned} \varepsilon &= y - \hat{y} \\ &= y - (\beta_0 + \beta_1 x). \end{aligned} \quad (7.4)$$

For a set of n years of data, the sum of the squares of these errors, SS_E , is minimised⁴⁷, i.e.:

$$\begin{aligned} \min SS_E &= \min \sum_{k=1}^n \varepsilon_k^2 \\ &= \min \sum_{k=1}^n [y_k - (\beta_0 + \beta_1 x_k)]^2. \end{aligned} \quad (7.5)$$

Equation (7.5) is minimised by setting its first partial derivatives to zero:

$$\begin{aligned} \frac{\partial SS_E}{\partial \beta_0} &= \frac{\partial \sum_{k=1}^n [y_k - (\beta_0 + \beta_1 x_k)]^2}{\partial \beta_0} = 0, \\ &= -2 \sum_{k=1}^n (y_k - \beta_0 - \beta_1 x_k) = 0 \end{aligned} \quad (7.6a)$$

and:

$$\begin{aligned} \frac{\partial SS_E}{\partial \beta_1} &= \frac{\partial \sum_{k=1}^n [y_k - (\beta_0 + \beta_1 x_k)]^2}{\partial \beta_1} = 0. \\ &= -2 \sum_{k=1}^n x_k (y_k - \beta_0 - \beta_1 x_k) = 0 \end{aligned} \quad (7.6b)$$

⁴⁶ In this chapter \hat{y} is referred to as “estimates” or “fitted values” when applied to cases within the training period (i.e. to cases used to estimate the regression parameters), and to “predictions” only when new values of x are applied. See sections 7.3.3.c and 7.3.3.d for a definition and discussion of the training period.

⁴⁷ The minimisation of the sum of the squared errors is by far the most commonly used form of estimation in seasonal climate prediction. The only other minimization criterion that has been used to any notable degree is that of minimising the sum of the absolute errors, and is known as “least absolute deviation” (LAD) regression. See Section 7.4.2.b for further discussion of LAD regression.

From Eq. (7.6), the two regression parameters can be obtained as:

$$b_1 = \frac{\sum_{k=1}^n [(x_k - \bar{x})(y_k - \bar{y})]}{\sum_{k=1}^n (x_k - \bar{x})^2} \quad (7.7a)$$

and

$$b_0 = \bar{y} - b_1 \bar{x}, \quad (7.7b)$$

where b_0 and b_1 are estimates of the parameters β_0 and β_1 , respectively.

The regression coefficient is closely related to Pearson's product moment correlation coefficient⁴⁸, r :

$$r = b_1 s_x s_y^{-1}, \quad (7.8)$$

where s_x and s_y are the standard deviations of x and y , respectively. The correlation coefficient is a widely used measure of the strength of linear association between the predictor and the predictand. Although it can be estimated using Eq. (7.8), it is more commonly calculated using:

$$r = \frac{\sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y})}{s_x s_y} \quad (7.9)$$

The numerator in Eq. (7.9) is related to the covariance by a factor of n , and will be positive if positive anomalies in both the predictor and the predictand tend to occur in corresponding cases, and will be negative if opposite anomalies tend to occur. Equation (7.9) defines the correlation as the standardised covariance. Frequently the correlation is squared, and it can then be interpreted as the proportion of the variance of the predictand that can be 'explained' using the predictor.

As an example, December–February 1961/62–2000/01 rainfall over Lusaka, Zambia, is shown as the y variable in Figure 7.1, and is regressed against the October value of the Niño3.4 index. Lusaka is located in part of southern Africa where El Niño (La Niña) conditions are frequently associated with below-normal (above-normal) rainfall. The correlation is -0.49, and is statistically significant at a 1% significance level, indicating that

⁴⁸ There are other correlation coefficients, but Pearson's is by far the most widely used, and unless specified otherwise, the term "correlation" refers to Pearson's correlation.

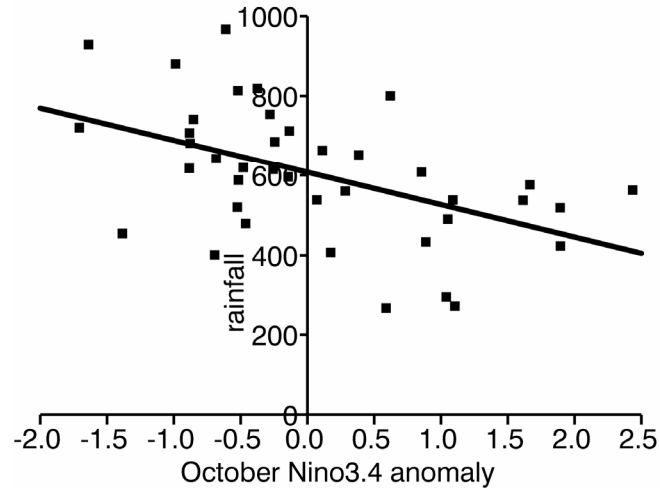


Figure 7.1 – Example of a linear regression model in which October values of the Niño3.4 index are used to predict December - February 1961/62–2000/01 rainfall totals for Lusaka, Zambia. The solid line represents the regression model.

there is a strong statistical basis for making a prediction. The figure shows that rainfall tends to decrease over Lusaka as the equatorial Pacific becomes warmer. The relationship with October values of the Niño3.4 index implies that a prediction can be made with a lead-time of one month using the formula:

$$\widehat{\text{rainfall}} = 607 - 81 \times \text{October Niño3.4}. \quad (7.10)$$

The negative regression coefficient in Eq. (7.9) means that the expected seasonal rainfall decreases by more than 80 mm for every 1°C increase in temperature in the central equatorial Pacific.

7.3.3.b Model formulation – multiple linear regression

When more than one predictor is used, a multiple regression model assumes the following form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \varepsilon. \quad (7.11)$$

Given m predictors and n cases (years of data), the regression model becomes:

$$\hat{y}_k = \beta_0 + \beta_1 x_{k,1} + \dots + \beta_m x_{k,m}. \quad (7.12)$$

Equation (7.12) has $p = m + 1$ parameters, and can be simplified in matrix notation to:

$$\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta}, \quad (7.13)$$

where \mathbf{X} is a $n \times p$ array in which the rows represent each year of data, and the columns represent each predictor, with the first column containing unity⁴⁹, and the $i + 1^{\text{th}}$ column containing the i^{th} predictor.

As with simple linear regression, the objective is to estimate the parameters $\boldsymbol{\beta}$ so that the sum of squares of errors is minimised:

$$\begin{aligned} \min SS_E &= \min \sum_{k=1}^n \varepsilon_k^2 \\ &= \min \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} \\ &= \min (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \min \mathbf{y}^T \mathbf{y} - 2\boldsymbol{\beta}^T \mathbf{X}^T \mathbf{y} + \boldsymbol{\beta}^T \mathbf{X}^T \mathbf{X} \boldsymbol{\beta} \end{aligned} \quad (7.14)$$

Similarly, Eq. (7.14) can be minimised by taking the first derivatives:

$$\frac{\partial SS_E}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^T \mathbf{y} + 2\mathbf{X}^T \mathbf{X} \boldsymbol{\beta} = \mathbf{0}, \quad (7.15)$$

which can be rearranged to give:

$$\boldsymbol{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}. \quad (7.16)$$

In practice, the inverse in Eq. (7.16) is difficult to calculate and can be prone to rounding errors if the predictors are inter-correlated, and so most statistical packages use alternative formulations and advanced linear algebra techniques, such as the singular value decomposition, to obtain the parameter estimates.

7.3.3.c Predictor selection

Unless the predictors to be used are predefined, the candidate predictors would normally be tested for inclusion in the final model that is to be used to make predictions. The standard approach is to include only those predictors in the final regression equation that contribute to a significant reduction in the size of the errors. Since the addition of *any* additional predictor into the model will always reduce the size of the errors, this reduction needs to be significantly large, i.e. the estimates need to improve sufficiently for us to be confident that the inclusion of the added predictor will effect an improvement in real-time predictions.

⁴⁹ This extra column is used for the regression constant, which is given as the first element of $\boldsymbol{\beta}$.

A commonly used procedure for selecting predictors is stepwise regression. There are three main forms of stepwise regression:

- *forward selection*: predictors are added one-by-one, with the remaining candidate predictor that reduces the size of the errors the most being added next, and continuing until the errors cannot be significantly reduced;
- *backward elimination*: all candidate predictors are initially included, and then predictors are removed one-by-one, with the predictor that increases the size of the errors the least being removed next, and continuing until the errors can only be increased significantly;
- *stepwise selection*: predictors are added one-by-one in the same way as for forward selection, but at each stage the included predictors are re-tested so that if the removal of any of these predictors results in an insignificant increase in the size of the errors they are removed.

All of these stepwise procedures require a criterion for deciding whether the change in the size of the errors is significantly large. The approach generally used is based upon the F -statistic, and involves a decomposition of the total sum of squares about the mean of the predictand, SS_T :

$$\begin{aligned} SS_T &= \sum_{i=1}^n y_i^2, \\ &= \mathbf{y}^T \mathbf{y} \end{aligned} \quad (7.17)$$

where y has been centred around zero by subtraction of the mean. The SS_T is decomposed into two components: the explained component as modelled by the regression model, SS_R , and the unexplained component or sum of the squares of the errors, SS_E , as defined in Eq. (7.14). The regression sum of squares is calculated as:

$$\begin{aligned} SS_R &= \sum_{i=1}^n \hat{y}_i^2 \\ &= \hat{\mathbf{y}}^T \hat{\mathbf{y}}, \\ &= (\mathbf{Xb})^T \mathbf{Xb} \\ &= \mathbf{b}^T \mathbf{X}^T \mathbf{Xb} \end{aligned} \quad (7.18)$$

so that $SS_T = SS_R + SS_E$. The F -statistic tests whether the change in SS_R for the predictor under consideration is significantly large compared to the mean of the squared errors, MS_E , after including the predictor. The MS_E is the SS_E divided by $n - p - 1$. Under the assumption that the predictor is unrelated to the predictand, the F -statistic is drawn from an F distribution with one and $n - p - 1$ degrees of freedom. A predefined value of this statistic

can be defined for a given level of significance (typically 0.05), and if the calculated F -statistic exceeds this value the predictor results in a significant improvement in the estimates of y . The procedure, however, is problematic, partly because of the sensitivity of the F -statistic to distributional assumptions (section 7.3.3), and because of problems related to multiplicity (section 7.4.1), which invalidate the significance tests.

Nevertheless, given the definitions of SS_T and SS_R in Eqs. (7.17) and (7.18), the ratio SS_R / SS_T provides an indication of the proportion of the total variability in the predictor that can be explained by the regression model. This proportion, denoted R^2 , is known as the coefficient of determination, and is the multivariate equivalent of the squared correlation coefficient (section 7.3.3). An adjusted R^2 is sometimes calculated to correct for the number of parameters in the model⁵⁰. The procedure described above based on the F -statistic is equivalent to selecting which of the two models (the one with and the one without the predictor under question) has the larger adjusted R^2 .

None of the stepwise procedures guarantees that the best possible set of predictors (i.e. the one that minimises the errors) is selected, and so one option is to search through all possible combinations to find that subset that reduces the size of the errors most significantly. Since this search can be computationally prohibitively expensive if the number of candidate predictors is large⁵¹, an alternative is to modify the simpler forward selection and backward elimination procedures described above by swapping out at each step any predictors that can effect an improvement in the model. The predictors are swapped one-by-one with the predictor that improves the model the most being introduced as replacement. The swapping continues until no further improvement is possible.

A somewhat different approach is to identify a model that makes a good set of independent predictions, as opposed to one that minimises the errors in estimating the data used to construct the model. The problem with minimising Eqs. (7.5) and (7.14) is that the model is optimised only to describe the relationship between the predictors and predictand over a set period, known as the training or calibration period (the period of the data used to construct the model), but there is no guarantee that this model will make good predictions when it is applied over a different period. Some procedures search for the set of predictors that make the best set of independent predictions by using only part of the data to construct the model and then examining the predictions for the data that was withheld. Techniques for

⁵⁰ Note that the adjusted R^2 cannot be interpreted as the proportion of variability explained.

⁵¹ Given k candidate predictors the number of possible combinations is $2^k - 1$.

performing this independent assessment are discussed in further detail in section 7.3.3.

7.3.3.d Model assumptions

Before assessing how well the regression model can predict the response variable, it is important to assess the validity of the model. If various assumptions about the data used in constructing the model cannot be upheld, the model parameters may be estimated incorrectly, and the predictions made in real-time will then be less accurate than expected. These assumptions are enumerated below. Alternative procedures for when these assumptions are invalid are discussed in section 7.4.

- Errors are identically and independently distributed (iid)

The forecast errors (Eq. 7.4) should show no tendency to increase or decrease in size either in the long-term or for identifiable sub-periods of the data. Similarly, the variance of the errors should not be related to values of the predictors (“homoscedasticity”). This latter restriction is often a problem when constructing statistical models to predict precipitation because forecast errors typically increase as the forecasted precipitation increases simply because there is a lower bound to precipitation.

In addition, the errors are assumed to be independent of each other. This assumption means that the model should show no tendency to underestimate or overestimate the observed values over a string of years. In combination with the assumption of a zero mean-error, the independence of errors means that each time a new prediction is made, the probability of overestimating (or underestimating) the observed value is 0.5 in all cases⁵². The Durbin-Watson test is recommended for testing independence of the errors, and works by identifying whether there is any autocorrelation in the errors (i.e. is it possible to “predict” the errors from previous errors?).

- Predictand is normally distributed

Although strictly it is only the model errors that need to be normally distributed, in practice, this distributional assumption about the errors is more often met when the predictand itself is normally distributed. In addition, if the predictand is not normally distributed, the regression parameters can be heavily influenced by the more extreme values. Since seasonal rainfall totals for many areas have a positively skewed distribution (see, for example, Figure 7.1), it is often advisable to transform the data so that the transformed data are normally distributed. Commonly used transformation

⁵² More generally, because of the assumption of fixed variance, the probability that the error will exceed any pre-defined value is a constant.

functions include the logarithm, and the square root and other power transformations.

- Linear relationship

If the relationship between (any of) the predictor(s) and the predictand is non-linear, Eqs. (7.2) and (7.11) are of the wrong “form”. The true form of the relationship(s) may be unknown, but more complex relationships can be examined using alternative regression models (section 7.4.2). Apart from testing for improvements in the predictions if a more complex model is used, it can be useful to reorder the predictions so that they are sorted by the value of (one of) the predictor(s) rather than chronologically, and then re-conducting the test for independence. If the true form of the relationship is quadratic, for example, but is assumed to be linear, the residuals will be of a similar sign at the beginning and end of the re-ordered series, and of the opposite sign in the middle.

- Uncorrelated predictors

For multiple regression, the model parameters can be estimated inaccurately when there are strong correlations between the predictors. The presence of strong correlations between predictors is known as multicollinearity, and is discussed in further detail in section 7.4.1.

7.3.3.e Model evaluation

Since measures of the errors in estimating the y values (“goodness of fit” measures), are as much a function of the number of parameters included in the model as they are of the quality of the model’s ability to describe the variability in the predictand, they are not necessarily very informative. In order to estimate how well the model can predict new values, a separate set of data that was not used to construct the model is required. Two approaches are used, and in both cases the data is divided into a “training” or “calibration” period, and an “independent” or “verification” period:

- *Cross-validation*: one year is withheld (together, optionally, with additional years immediately preceding and succeeding; this omitted period is known as the cross-validation window), and the remaining years are used to train the model. A prediction is made for the omitted year or the year in the middle of a window larger than one, and the procedure is repeated until a prediction has been made for each year (Figure 7.2a and b).
- *Retroactive validation*: the model is trained using only the first few years of the data, and a prediction is made for the year immediately after the end of the training period. The model is then updated, adding the year just predicted to the training period, and a prediction for the follow-

ing year is made (Figure 7.2c). This procedure is continued until a prediction for the last year has been made. (Sometimes the subsequent k years are predicted, where $k > 1$, and the model is only updated every k years).

a Leave-one-out cross-validation

1951	Predict 1951	Training Period			
1952	Training period	Predict 1952	Training Period		
1953	Training period		Predict 1953	Training Period	
1954	Training Period			Predict 1954	Training period
...	Training period				Verification Period
2000	Training Period				Predict 2000

b Leave-three-out cross-validation

1951	Predict 1951	Omit 1952	Training Period		
1952	Omit 1951	Predict 1952	Omit 1953	Training Period	
1953	Training period	Omit 1952	Predict 1953	Omit 1954	Training Period
1954	Training Period		Omit 1953	Predict 1954	Omit 1955
...	Training period			Omit	Verification Period
2000	Training Period				Predict 2000

c Retroactive validation

1981	Training period 1951-1980	Predict 1951	Omit 1982-2000		
1982	Training period 1951-1981		Predict 1952	Omit 1983-2000	
1983	Training period 1951-1982			Predict 1983	Omit 1984-2000
...	Training period				Verification Period
2000	Training Period 1951-1999				Predict 2000

Figure 7.2 – Schematic diagrams illustrating the procedure for (a) leave-one-out cross-validation, (b) leave-three-out cross-validation, and (c) retroactive validation.

In each case, the objective is to generate a set of “out-of-sample” predictions. These predictions need to be independent of the data used in the training set, but assuring complete independence is exceptionally difficult, particularly with cross-validation. One of the main ways in which “leakage” of information from the training to the verification sample is allowed to occur is through a failure to reselect the predictors adequately at each step. It is important that the predictors are allowed to be reselected rather than only allowing the model’s parameters to be recalculated⁵³. Ideally each training period should be independent of each other, but since that is impractical because of limited sample sizes, some effort to ensure that at least some of the training periods differ should be made. In cross-validation this independence can only be achieved by using a fairly large window.

Retroactive validation closely mimics the operational generation of predictions, and so should give a realistic estimate of how well the model would have performed if it had been operational since the first year of the independent predictions (although selection of candidate predictors by using all the data can bias the results). The downside of retroactive validation is that predictions are made only for a subset of the data, and so the small sample size will contribute to large errors in the estimates of the quality of the predictions.

In cases where the predictor(s) is (are) specified and the distributional assumptions described in the previous section do not hold, bootstrapping of the model parameters should be conducted. Bootstrapping involves randomly re-sampling pairs of predictor and predictand values, and then recalculating the regression using the resample. There are many ways of designing a bootstrap procedure, but the standard approach is to generate a sample that has the same number of cases as the original sample. The cases are drawn with replacement, for otherwise the bootstrap sample would be identical to the original sample. A large number of bootstrap samples are generated, and regression models constructed for each one. The distribution of the regression parameters provides an indication of the uncertainty in estimating the “correct” parameters.⁵⁴

⁵³ By reselecting the predictors at each step it is quite possible that the actual set of predictors that are used to make an operational forecast are not actually selected in some or even any of the cross-validation steps. This failure to test using the operational predictors may seem problematic, but an essential part of the cross-validation procedure is to test the predictor selection process.

⁵⁴ Although not widely performed, one way of estimating the uncertainty in a prediction would be to make a suite of predictions using models constructed using the bootstrap samples. More widely used methods are discussed in the following section.

7.3.3.f Scoring metrics

Given a set of independent predictions, the most commonly used metric to calculate how well these predictions match the observed outcomes is the correlation coefficient. The correlation coefficient was introduced in section 7.3.3, where it was used to measure the strength of the linear association between the predictor(s) and predictand. To use the correlation for forecast verification, simply replace x with \hat{y} in Eq. (7.9). Note, however, that the correlation is not a measure of forecast accuracy for two reasons: the subtraction of the means of x and y in the numerator eliminates any bias in the forecasts, and the division by the respective standard deviations eliminates any variance bias. (See section 7.3.3 and **Chapter 8** for definitions of accuracy, bias, and variance bias.) As a result, predictions of rainfall, for example, that are consistently too wet or too dry, and vary too much or too little can still achieve a perfect verification score. In the context of statistical models, such problems are not usually very severe because the predictions should be reasonably well calibrated over the training period. As a result, the mean bias should be fairly small, although in most cases the variance will be underestimated, simply because in an imperfect model predictions err towards the climatological mean.

The squared correlation coefficient is often quoted as the percentage of variance of the observed values that is successfully predicted. While technically correct, this percentage is often misinterpreted as some measure of how frequently the forecasts are “correct”. In the context of the deterministic predictions from regression models, “accuracy” is a more appropriate quality of the forecasts than correctness because the predicted and observed values will always differ if only by a very small amount, and so the predictions are never “correct” in a strict sense. Accuracy generally is indicated using an average of some measure of the errors. The mean squared error, introduced in section 7.3.3, is a natural choice because it is a quantity that has been minimised when the model was constructed, but is not particularly intuitive otherwise. The root mean squared error resolves the conceptual problem of interpreting squared errors, but the mean absolute error is the simplest to understand: it indicates by how much, on average, the predictions differ from the observed outcomes. A still more informative approach would be to indicate in a contingency table or histogram how frequently errors of different magnitude occur.

Other widely used metrics are based on the contingency table: it has become popular to assign the observed values to one of three equiprobable categories, labelled “below-normal”, “normal”, and “above-normal”, with “below-normal” referring to the driest/coldest third of cases, and the other

categories defined accordingly⁵⁵. The deterministic forecasts can be classified into one of these three categories, and a table comparing the forecast and observed categories can then be constructed. An example is shown in Table 7.1a for 30 years of cross-validated predictions of December–February Lusaka rainfall using only the Niño3.4 index as predictor. The “correct” predictions are shown in the diagonal cells from top left to bottom right.

a)

		Predictions			Total
		A	N	B	
Observations	A	3	7	0	10
	N	0	7	3	10
	B	1	5	4	10
Total		4	19	7	30

b)

		Predictions			Total
		A	N	B	
Observations	A	5	4	1	10
	N	2	4	4	10
	B	3	2	5	10
Total		10	10	10	30

Table 7.1 - (a) Contingency table and (b) variance-adjusted contingency table of cross-validated predictions of December - February 1971/72–2000/01 rainfall totals for Lusaka, Zambia, using the October Niño3.4 index as sole predictor. The categories are equiprobable, and are marked B for below-normal, N for normal, and A for above-normal.

There is a wide range of summary measures of such contingency tables, but they are not discussed here because the loss of information as a result of the categorization of the observations and predictions, and deterministic nature of the predictions mean that such an interpretation of the climate prediction information is undesirable. The interested reader is referred to Jolliffe and Stephenson (2003) and Wilks (2005) for details.

⁵⁵ Categories do not have to be equiprobable, and more (or less) than three categories can be defined. The principles of verification remain the same, however.

The number of predictions of the normal category is higher than for the other categories because of the lower variance of the predictions compared to the observations. As a result, the variance of the forecasts is sometimes increased artificially so that the number of predictions in each category is equal. The resulting contingency table is shown in Table 7.1*b*. Such variance adjustment is problematic because the squared errors are no longer minimised, and it can be seen from Table 7.1*b* that there is no improvement in the total number of correct predictions (5+4+5 compared with 3+7+4), while there is an increase in the number of two-category misses (i.e. predictions of above-normal when below-normal occurred, or vice versa). Variance-adjustment should therefore be discouraged.

Ideally, if the forecasts are categorised they should be expressed as probabilities. Methods for generating probabilistic forecasts from the deterministic predictions of regression models are discussed in the following section. The verification of probabilistic forecasts is a complex issue, and is discussed in detail in [Chapter 10](#).

7.3.3.g *Generating probabilistic forecasts*

Once the regression model has been constructed, predictions can be made using Eqs. (7.3) and (7.13) given new values of the predictor(s). However, these equations give only a “best-guess” of the outcome, and no indication of the uncertainty is provided. There are a number of ways in which this best-guess forecast can be converted to a probabilistic forecast, but the most reliable procedure is to use information about the variance of the errors in estimating previous known values. The error variance is widely used to define a prediction interval on the forecast, although it is possible to obtain probabilities for predefined categories as well. If the errors in the forecasts are assumed to be Gaussian, these probabilities can be calculated by integration of the *t*-distribution using the best-guess as the mean and the error variance as the variance. (See [Chapter 8 section 8.5.1](#) for discussions on different ways of communicating forecast uncertainty.) The error variance is normally calculated from the fitted values, although the errors in the cross-validated forecasts could be used instead, and may be more reliable.

Alternative approaches include using contingency tables that compare the category of the forecast with the observed category for a set of forecasts. Then if 60% of the times that the forecast has indicated below-normal rainfall the observation was also below-normal, for example, the forecast would specify a 60% probability of below-normal rainfall the next time the forecast indicates below-normal. There are two problems with this approach: very large samples are required to estimate the probabilities reliably, and; no distinction is made between the probabilities issued when the forecast indicates well below-normal rainfall, and when it indicates marginally be-

low-normal. The large differences in the amount of rainfall that can be classified as “below-normal”, for example, could be offset by increasing the number of categories, but only at the cost of requiring still larger samples. Given these problems, the use of contingency tables to obtain forecast probabilities is not recommended. Instead there is a suite of statistical procedures that can be used to obtain these probabilities directly rather than estimating a best-guess and then trying to account for the uncertainty subsequently. These procedures are discussed in section 7.4.2.

7.4 Alternative statistical methods to linear regression

Linear regression forms the basis for a number of more sophisticated statistical techniques that have been used in seasonal climate prediction. Some of these techniques are discussed in section 7.4.2, all of which have in common an attempt to estimate a “best-guess” forecast. Some alternative statistical techniques that estimate forecast probabilities without providing a best-guess are considered in section 7.4.2. However, to understand the motivation for using any of these methods, it is first helpful to consider some of the limitations and potential pitfalls of linear regression, and these issues are outlined in section 7.4.1.

7.4.1 PROBLEMS WITH LINEAR REGRESSION

The problems and potential pitfalls listed in this section are not exclusive to linear regression, but are listed to provide a context for understanding the more sophisticated techniques described in sections 7.4.2 and 7.4.3. In many cases the alternative techniques attempt to address only a subset of the problems listed below.

7.4.1.a Multiplicity

One of the primary difficulties in using linear regression for seasonal climate forecasting is identifying the predictors to use in the model. Most frequently, predictors used are measurements of SSTs, but land-surface characteristics and atmospheric indices are also used for forecasting in countries such as India where the use of such variables has been supported by extensive research on seasonal predictability. Whether or not SSTs are used exclusively, the pool of candidate predictors is vast, and the problem arises of which subset of these predictors should be included in the regression model. The temptation is to choose the predictors that are best correlated with the predictands, but the probability of identifying highly, but

spuriously, correlated predictors increases⁵⁶ as the pool of candidate predictors is expanded. This problem is known as “multiplicity”, and the search for predictors by repeated testing of the strength of statistical relationships is known as “fishing”, and almost invariably results in the creation of a statistical model that performs worse than anticipated when used operationally.

One reason why “fishing” results in models that perform poorly in operations is that standard tests of statistical significance used in constructing a statistical model assume that the predictors to be used in the regression model have already been selected, and these tests become invalid when only the models that give the best results are selected. If a number of regression models are tested with the aim of identifying those that work well, then problems of multiplicity arise. Standard significance tests require adjustment for multiplicity, otherwise there is an increased danger of accepting predictors that should not be included in the model, and/or of overestimating the strength of the model’s predictive capability. This selection of spurious, or of too many, predictors is known as “over-fitting”.

Cross-validation (section 7.3.3) is used to test for over-fitting. Leave-one-out cross-validation appears to be the standard approach in the atmospheric sciences (leave-*k*-out is used if the data are autocorrelated, but *k* typically is set only to a maximum of twice the decorrelation time). However, it is not widely recognised in the atmospheric sciences literature that a substantial proportion of the data needs to be omitted to obtain unbiased results. How much data should be omitted remains a question for further research, but there have been suggestions that it should be as much as 40–60% (Xu and Liang 2001). Frequently, therefore, the problems of multiplicity are not adequately addressed.

An aspect of multiplicity is evident not just when constructing a model with a large pool of candidate predictors, but also when constructing a number of models, perhaps for different stations and/or seasons. If numerous models are constructed, the probability of finding at least one that gives spuriously “good” predictions increases, and so the statistical significance of the overall set of results needs to be assessed. Tests for “field significance” have been designed to address this question. Multiplicity problems can apply to GCM forecasts as well since forecasts are made for a large number of locations, variables, lags, and target periods.

7.4.1.b Multicollinearity

Multicollinearity is a problem that sometimes arises when more than one predictor is used in a regression model. If the predictors used are themselves highly correlated with each other, errors in estimating the model parameters

⁵⁶ I.e. the probability of making a type-I error increases.

can become substantial. The errors in these parameter estimates can give poor predictions when new values of the predictors are applied to the model, and can also create problems in interpreting the regression coefficients. Whereas multiplicity results in bad forecasts because of the inclusion of incorrect predictors in the model, multicollinearity can cause bad forecasts even when the correct predictors are included simply because the regression parameters may be poorly estimated.

To illustrate the difficulty of interpreting regression parameters when predictors are correlated, consider a simple multiple regression model to predict the March values of the NiNO3.4 index from the January and February values. Using data for 1971–2000, the regression coefficients for January and February, respectively are -0.395 and 1.216, which seems to imply that the March value is negatively correlated with the January value, whereas one would expect a slightly weaker positive correlation than for February. However, when the January and February values are used in separate models as the only predictors, the coefficients change to 0.628 and 0.761, respectively, showing that the values in both months are positively correlated with that in March.

7.4.1.c Non-linearity

Linear regression assumes a linear relationship between the predictor(s) and the predictand. This assumption means that for a given change in the value of a predictor, (e.g. a 1°C increase in SST in a specified area), the expected change in the predictand (e.g. an increase in seasonal rainfall of 100 mm) is the same regardless of the actual sea temperature, and regardless of the values of the other predictors. Given the non-linear nature of the atmosphere the linearity assumption seems inherently unreasonable, and the flexibility to model non-linear relationships statistically may be desirable. In practice, however, the linearity assumption is often a reasonable approximation, and even where it is not, the degrees of freedom required to identify the correct form of the relationship are likely to be lacking.

7.4.1.d Assumptions about data distribution

In addition to the linearity assumption, linear regression assumes that the predictand (but not necessarily the predictors) is normally distributed (see [Chapter 9](#)). While this assumption may be quite reasonable for variables such as geopotential heights, for other variables the data may not be normally distributed, and fitting a linear regression then becomes problematic. Although the distribution of surface air temperatures is skewed, this can generally be ignored because the skewness is not usually severe. However, with precipitation, skewness can be marked (see examples in [Chapter 8, section 8.3.1](#)), and there is the related problem that precipitation has a lower

limit of zero. It makes no sense for a regression line on precipitation to extend below zero since negative precipitation is meaningless, but a linear regression model is unaware of such a constraint. The lower limit on precipitation also means that even if a regression model is fitted, the errors are usually larger for larger precipitation rates than for rates close to zero. This increase in the variance of the errors in estimating precipitation for larger precipitation amounts violates the homoscedasticity assumption of multiple linear regression. Although these problems could be addressed by using certain forms of generalised linear models (see section 7.4.3), they are frequently ignored, or assumed not to be problematic.

7.4.2 REGRESSION-BASED STATISTICAL PREDICTION TECHNIQUES

7.4.2.a Power and non-linear regression

Even when the relationship between the predictor and predictand is non-linear, a transformation of the values of the predictors and/or predictand may make it possible to treat the problem as linear. The most commonly used transformations are power transformations (e.g. using the square or the square-root of the predictors), and adding these to the pool of predictors. The resulting models, known as power regression models⁵⁷, have been used extensively in statistical predictions of the Indian monsoon, for example. However, caution has to be taken since the problem of multiplicity is exacerbated by expanding the number of candidate predictors, and theoretical justifications for the power transformations should be supplied. In addition, multicollinearity is introduced with most power transformations. Power regression is sometimes used in seasonal forecasts of climate impacts, where non-linear relationships between climate variables and the application data in question have a theoretical basis (e.g. Chapters 12 and 13). Other examples of non-linear regression include exponential models, which are used more frequently in forecasting impacts than in forecasts of seasonal climate per se.

Compared to power regression models, neural networks constitute a considerable increase in the complexity with which nonlinear relationship can be modelled. Neural networks are a recent development in seasonal climate prediction, but have been applied successfully, and have been implemented as the statistical atmospheric component in hybrid coupled models. The neural networks are constructed by optimizing sets of weights applied to the predictors, which are then transformed using a non-linear function (usually the hyperbolic tangent), and then further weighting func-

⁵⁷ Polynomial regression models are special cases of power regression, allowing only integer powers to be used.

tions are applied to provide estimates of the predictand values. The weights are optimised so that the squared errors in the estimates are minimised, as with linear regression. Because of the large numbers of model parameters involved, care has to be taken to avoid over-fitting.

7.4.2.b Regression models for non-normally distributed data

Although linear regression assumes that the data being analysed are normally distributed, the procedure can be generalised to allow for predictands with alternative distributions. These generalised linear models (GLMs) are discussed in more detail in section 7.4.3, where versions of GLMs for estimating probabilities are considered. However, there are forms of GLMs for data with a Poisson distribution that are suitable for modelling data that are recorded as counts, and these have been applied in seasonal forecasting of tropical cyclones. Versions are also available for data with a gamma distribution that would be suitable for forecasts of rainfall, but these have not been widely used.

A primary reason why linear regression becomes problematic when the predictands are not normally distributed is that the more extreme observations (for example the very wet years) have an undue influence on the regression parameters. While GLMs address this problem by making it possible to assume distributions that are more representative of the data, another alternative is to use regression models that are less sensitive to extreme values. There are two ways in which this sensitivity can be reduced. In robust regression, one option is to reset all errors (i.e. squared differences between the observed and the fitted value) exceeding a maximum value to this threshold. The procedure is not widely used. The second approach is to redefine how the errors are calculated: specifically, instead of squaring the errors, which tends to exaggerate the magnitude of large errors, the absolute errors can be used. This procedure is known as least absolute deviation (LAD) regression, and has been used in tropical cyclone forecasting, for example.

7.4.2.c Ridge regression

Ridge regression constitutes an attempt to address the problem of multicollinearity by placing constraints on the model parameters. In effect the procedure artificially inflates the variances of the predictors relative to their covariances, and thus underplays the effects of the inter-correlations when estimating the model regression coefficients. Ridging is used in the constructed analogue procedure, in which a least squares estimate of the spatial pattern of the most recently observed values of the predictands is obtained by weighting the patterns for all years in the historical data. As an example of a simple constructed analogue model, consider the problem of forecast-

ing the December Niño3.4 index from the June value. Assume that the June and December values of the index are known for 1971–2000, and that the June 2001 value is available to make a forecast for December 2001. Weights would be assigned to the June values for 1971–2000 to estimate the June value for 2001. These same weights would then be applied to the December 1971–2000 values to construct a forecast for December 2001. Given that the number of weights to be calculated (30) is larger than the number of values being estimated (1), there is no unique solution to the weights, but the *ridging* helps to provide a stable solution.

7.4.2.d Principal components regression

Principal components regression (PCR) improves on ridge regression by addressing some of the problems arising from both multicollinearity and multiplicity. The only difference between multiple regression and PCR is that in PCR the principal components of the predictors are used in the model instead of the original predictors themselves. Principal components are optimal summaries of large sets of data, obtained by defining sets of weights, or “loadings”, that are applied to obtain a linear combination of the original data. They are ideally suited to the problem at hand, since they will reduce a large candidate pool of predictors to a much smaller number, while retaining much of the information in the original data. In addition, each of the principal components is uncorrelated with all the others, and so problems of multicollinearity are avoided. More complex versions of principal component analysis can be used in PCR that represent, for example, modes of variability that have an evolutionary component, and are discussed further in the next section.

In theory it is possible to expand a PCR equation into an equivalent multiple regression equation given the PCR coefficients and the loadings used to define the principal components. The coefficients of this expanded multiple regression have smaller error variance than if the coefficients had been estimated directly, because the negative effects of multicollinearity are usually associated with the higher order principal components that would generally be omitted from the analysis. However, the coefficients are biased, and so problems of interpretation remain. Despite these issues, and problems in determining the number of principal components to retain in the model, principal components regression is an attractive alternative to multiple linear regression.

7.4.2.e Maximum covariance analysis, canonical correlation analysis, and redundancy analysis

When making predictions for a number of different stations or gridpoints, principal components regression can be an inefficient procedure since sepa-

rate models have to be constructed and tested for each location. In addition, if the predictands are inter-correlated, it is possible for predictions at one or more of the locations to be somewhat inconsistent with those at others because of different sampling errors in the estimated regression coefficients, or even in the selection of predictors, for models at neighbouring sites. There are various techniques that can be used to make predictions at a set of locations. These techniques include canonical correlation analysis (CCA), redundancy analysis, and maximum covariance analysis⁵⁸ (MCA). These techniques are widely used in spatial downscaling problems (Chapter 8).

The basic principle behind all of these techniques involves forecasting modes or spatial patterns of variability spanning across the region of interest rather than making forecasts for individual locations. In this context, a mode is akin to a weighted average⁵⁹ of the individual locations. More than one mode can be predicted, and the predictions for these modes are then superimposed to construct forecasts for all locations. The modes are predicted using a second set of modes obtained from the predictors so that spatial patterns of variability in the predictors are used to predict spatial patterns in the predictands. If \mathbf{U}_X and \mathbf{U}_Y are the weights for the predictors and predictands, respectively, the modes, or new variables, are:

$$\mathbf{Z}_X = \mathbf{X}\mathbf{U}_X, \quad (7.19a)$$

$$\mathbf{Z}_Y = \mathbf{Y}\mathbf{U}_Y. \quad (7.19b)$$

Maps of the weights are frequently plotted to indicate the coupled spatial patterns. As an example, the first coupled mode (obtained using CCA) of September SSTs for the Indian Ocean and October–December precipitation over part of East Africa is shown in Figure 7.3. The mode suggests that warming in the western tropical Indian Ocean with cooling in the eastern tropical Indian Ocean and far western Pacific (Figure 7.3a) can be used to predict anomalously wet conditions over the bulk of Tanzania and Kenya (Figure 7.3b). The opposite precipitation pattern would be predicted given a reversal of the anomalous zonal temperature gradient in the tropical Indian Ocean. The temporal variability of these modes is shown in Figure 7.3c; the correlation between the modes is 0.706.

⁵⁸ Maximum covariance analysis is frequently referred to as singular value decomposition (SVD) or SVD analysis. This nomenclature, however, is confusing because SVD is often used to perform other analyses, including multiple regression, principal components analysis, and CCA. Von Storch and Zwiers (1999) propose calling the technique maximum covariance analysis.

⁵⁹ More strictly, because the sum of the squares of the weights rather than the sum of the weights per se, is required to be unity, the modes are a “weighted sum” or a “linear combination”.

The differences between MCA, CCA, and redundancy analysis are in the properties of the weights that define the modes:

- in MCA each pair of modes has maximum covariance;
- in CCA each pair of modes has maximum correlation;
- in redundancy analysis the explained variance in the predictand modes is maximised.

(Compare principal component analysis, in which the aim is to define a set of weights for either the predictors or the predictands that generate new variables with maximum variance.) For MCA, the covariance between the modes is:

$$\mathbf{C} = \mathbf{Z}_X^T \mathbf{Z}_Y. \quad (7.20)$$

The covariance matrix \mathbf{C} is a diagonal matrix with the diagonal elements defining the covariances of the coupled modes of predictors and predictands. Equation (7.20) can be written in terms of \mathbf{X} and \mathbf{Y} by substituting from Eq. (7.19):

$$\begin{aligned} \mathbf{C} &= (\mathbf{X}\mathbf{U}_X)^T \mathbf{Y}\mathbf{U}_Y \\ &= \mathbf{U}_X^T \mathbf{X}^T \mathbf{Y}\mathbf{U}_Y \end{aligned} \quad (7.21)$$

$\mathbf{X}^T \mathbf{Y}$ is the covariance of \mathbf{X} and \mathbf{Y} (i.e. the covariance matrix of the original predictors and predictands) and so Eq. (7.21) can be rearranged to express this covariance matrix, \mathbf{C}_{XY} , in terms of the diagonal matrix \mathbf{C} , and two orthogonal matrices:

$$\mathbf{C}_{XY} = \mathbf{U}_X^T \mathbf{C} \mathbf{U}_Y. \quad (7.22)$$

In other words, the weights \mathbf{U}_X and \mathbf{U}_Y that maximise the covariances between the spatial modes of predictors and predictands can be obtained from a singular value decomposition of the covariance matrix of the original predictors and predictands. Then, given a new set of predictors, \mathbf{x} , forecasts, $\hat{\mathbf{y}}$, can be generated:

$$\hat{\mathbf{y}} = \mathbf{x}\mathbf{U}_X \Sigma_X^{-1} \mathbf{C}\mathbf{U}_Y^T, \quad (7.23)$$

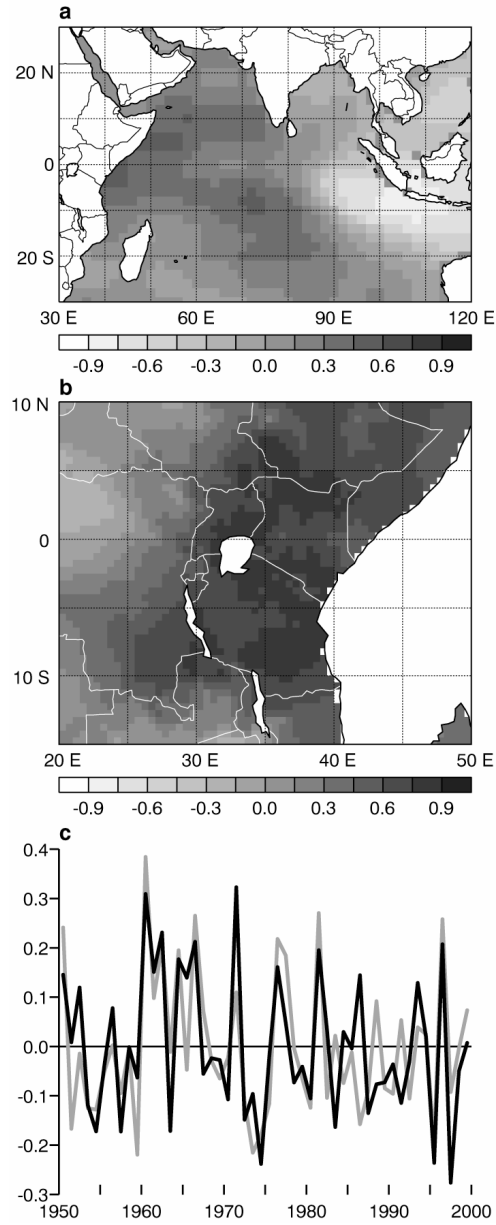


Figure 7.3 – Example of the first coupled mode of (a) September 1951–2000 sea surface temperatures for part of the Indian Ocean used to predict (b) October–December 1951–2000 precipitation over East Africa. Both datasets were pre-filtered by using only the first few principal components. The maps show the correlations between the original gridded data and the respective temporal scores (c) for the predictor (black) and predictand (grey) components of the first canonical coupled mode.

where Σ_X is a diagonal matrix containing the variances of the Z_X . Only those coupled modes that explain a large proportion of the total variance are used in the prediction, and so typically only the first few coupled modes are retained. Effectively, the smaller diagonal elements of the matrix C effectively are set to zero.

However, Eq. (7.23) does not provide least-squares estimates of the predictands, and so MCA is not regularly used in seasonal climate forecasting. Instead MCA is more useful in identifying coupled modes of, for example, SST fields and rainfall that may provide a basis for seasonal forecasting. A much more commonly used variant of MCA in prediction problems is CCA, which aims to identify alternative sets of weights, V_X and V_Y ⁶⁰, that maximise the correlations rather than the covariances between the modes of variability. In CCA the modes defined in Eq. (7.19) are first standardised, replacing U_X and U_Y by V_X and V_Y , respectively, so that C in Eq. (7.20) becomes a squared correlation matrix, R . Predictions, given a new set of predictors, are then given by:

$$\hat{y} = xV_XRV_Y^{-1}. \quad (7.24)$$

In practical terms, CCA identifies linear combinations of predictors that can successfully predict linear combinations of the predictands, regardless of how much of the total variance either linear combination explains. Consequently, there is a danger of identifying well-correlated modes of variability that do not explain much of the total variability. Although the objective in MCA of maximising the covariances rather than the correlations between the modes may seem more pertinent, MCA is also problematic in that the covariances are maximised in part by the variances of the modes for the predictors, and so it is possible that the total explained variance of the predictands is low. Further, both methods are subject to interpretation problems, and neither approach is likely to identify robust and easily interpretable modes of variability. Redundancy analysis is a third option that deserves further attention. Redundancy analysis replaces Z_X in Eq. (7.19a) with the standardised values, and thus seeks to maximise the explained variance in the predictands without necessarily using the largest modes of the variability in the predictors. Redundancy analysis can thus be seen as intermediate between CCA and MCA. In practice, differences in the results of the various techniques are usually minimal.

In most applications of MCA and CCA in the climate literature, the observations and forecasts are pre-filtered by using a subset of the principal components of the data. While the pre-filtering simplifies the solution of the CCA or MCA, the computational gain is lost through having to calculate

⁶⁰ Note that V_X and V_Y are not orthogonal matrices, whereas U_X and U_Y are.

the principal components. Instead, the main advantage of the pre-filtering is that the noise levels in both the forecasts and the observations are reduced, and so the chances of finding spurious relationships are decreased. This advantage is likely to be greater for CCA than for MCA because the former does not require the coupled modes to represent large proportions of the total variance of the original data.

7.4.2.f Other principal component analysis-related techniques

As mentioned in section 7.4.2.d, principal components can be useful as predictors. There is a hierarchy of sophisticated ways in which these components can be defined. In the simplest formulation, the principal components are defined using a set of predictor variables all of which represent measurements synchronous with each other. Prediction using principal components of SSTs at various locations, but all measured at the same time, would be an example. This form of principal components regression is discussed in section 7.4.2.d.

If the predictors are measured at a number of different lags, the principal components become “extended” empirical orthogonal functions (EOFs)⁶¹, whose computation is equivalent to that of multi-channel singular spectrum analysis. For example, SSTs for a set of locations measured at a number of different times of the year are sometimes used to predict future SSTs. If a single predictor is used in this context so that the principal components are calculated only from the autocorrelation (or auto-covariance) of this series, the technique is known as singular spectrum analysis (SSA). Although SSA has not been used widely in seasonal climate forecasting, it has been used in an attempt to identify the predictable component of the Indian monsoon variability. Similarly, complex EOFs have been used in predictability studies, but have not been widely applied in seasonal climate forecasting. Complex EOF analysis, sometimes called Hilbert singular decomposition, involves advancing all oscillatory components of any wavelength in the data by 90°, and including these as imaginary components in a principal component analysis. The procedure allows lags to be identified in modes of variability.

Principal oscillation pattern (POP) analysis is fundamentally different to the techniques described above. It performs an eigenvalue decomposition of the matrix of first order autoregressive (AR-1) coefficients, and hence identifies optimal multivariate AR-1 models that can be used for prediction

⁶¹ Empirical orthogonal functions are the loadings that define the principal components. Although some authors have drawn a distinction between principal component analysis and empirical orthogonal function analysis based on the normalization of the eigenvectors (Richman 1986), this distinction is not widely adhered to and the two are in most cases synonymous (von Storch and Zwiers 1999; Joliffe 2007).

purposes. POP analysis has similar objectives to complex EOF analysis in seeking to identify evolutionary modes of variability, but has been more widely used than the latter in seasonal prediction. Linear inverse modelling is a version of POP analysis.

7.4.2.g Autoregressive models and optimal climate normals

Linear inverse modelling and POP analysis are sophisticated versions of simpler models known as autoregressive models. Autoregressive models are mathematically the same as linear regression models except that the predictors are the same variable as the predictand, only measured at different lags. So, for example, if the Niño3.4 index is forecasted with a regression model using only earlier values of the index, then this model would be autoregressive. The best known example of such a model is the CLIPER (CLImatology and PERsistence) model that has been used to forecast the ENSO phase using lagged and autoregressive relationships. The basic principle involved is that some variables, such as SSTs, change slowly, and so recent evolution can be used as a guide to future values. The name CLIPER implies that future values are predicted using a combination of: the seasonal mean value (climatology) towards which the value of the predictand is expected to drift at increasingly long lead-times, and; the most recently observed anomalies, that are expected to decay⁶² only slowly (persist).

A special case of using persistence and climatology as a forecast is that of optimal climate normals (OCNs). In most cases of seasonal climate forecasting, a forecast is made by projecting the most recently observed climate state into the future, i.e. from the previous day (or month or perhaps season) into a coming season. However, with OCN a forecast is made under the assumption that a good guide to the climate conditions for the target season are the conditions that have been observed for the same season over the last few years. The forecast for the coming season is then simply the average of the last few years, and the objective is to identify the number of years to average to give the best forecast. The idea is that the 30-year standard climatological period can be improved upon in some cases when there is low-frequency variability (e.g. inter-decadal variability or trend) in the climate. Using OCNs is sometimes a useful option in areas with inherently low seasonal predictability.

⁶² It is possible, such as when forecasting ENSO anomalies at certain times of the year, for anomalies to grow in a CLIPER model (Knaff and Landsea 1997), but such cases are unusual.

7.4.3 PROBABILISTIC STATISTICAL PREDICTION TECHNIQUES

Rather than trying to estimate a best-guess forecast value and then accounting for the uncertainty in this forecast, there are a number of statistical techniques that can be used to estimate forecast probabilities directly. Some of these methods are alternative versions of the regression models mentioned in section 7.4.2, and are described in further detail in section 7.4.3.a, while others are based on classification problems, and are discussed in section 7.4.3.b. In section 7.4.3.a statistical procedures that are similar to ensemble forecasting are described.

7.4.3.a Generalised linear models

Although multiple regression can be used to estimate probabilities as the dependent variable, this is not generally advised because there is no constraint that the estimated probability is between zero and one, and because the distributional assumptions of the procedure are violated (Wilks 2005). Instead a variety of models that are ultimately based on linear regression are available. Although these generalised linear models are closely related to linear regression they are discussed separately in this section.

Generalised linear models are based on the standard linear regression equation:

$$\eta = \boldsymbol{\beta}^T \mathbf{x}, \quad (7.25)$$

where $\boldsymbol{\beta}$ is the set of regression parameters, and \mathbf{x} is the set of predictors. The linear predictor η is related to the predictand, which in this case is a Bernoulli variable with mean \hat{p} , via a link function. The three most commonly used link functions for Bernoulli variables are:

$$\eta = \log \left[\frac{\hat{p}}{1 - \hat{p}} \right], \quad (7.26a)$$

$$\eta = \Phi^{-1} [\hat{p}], \quad (7.26b)$$

$$\eta = \log [-\log [1 - \hat{p}]], \quad (7.26c)$$

where Φ^{-1} is the inverse normal distribution function. These link functions are known as the logit, probit, and complementary log-log functions, respectively. In practice, the differences between the three are minimal, but the logistic link is the most widely used, and easiest to compute.

Instead of training the model using observed rainfall or temperatures, for example, the predictand has to be categorised into one of two groups. For example, in Figure 7.1a December 1950–2000 values of the Niño3.4

index are shown as anomalies and plotted against the June values. The regression line and the scatter of values imply a reasonably strong relationship between the phase of ENSO in June and that six months later. In Figure 7.4b, all the values of the December Niño3.4 index that exceed the upper quartile are converted to a value of 1, and all the values less than the upper quartile to a value of 0. The values on the x -axis (the June Niño3.4 index) are left unchanged. Rather than trying to fit a straight line to the data points, an S -shaped curve is used. Eqs. (7.26a-c) are different ways of converting a straight line to an S -shaped curve that ranges between 0 as a minimum, and 1 as a maximum.

In this example of a generalised linear model, observations are listed either as 0s and 1s, and the fitted curve is interpreted as providing an estimate of the probability that future values will exceed the threshold used to define the categories (i.e. the probability that the December Niño3.4 index will exceed the upper quartile). The limitation to only two categories can be too restrictive, but it is possible to further divide the categories either by nesting models, or by simultaneous fitting of parallel models.

The forms of generalised linear models described above, resolve issues related to data distribution assumptions, of indicating forecast uncertainty, and, to some extent, that of linearity, but do not address the problems of multiplicity and multicollinearity. The latter two problems can be addressed in similar ways to that for linear regression, e.g. by using principal components as predictors.

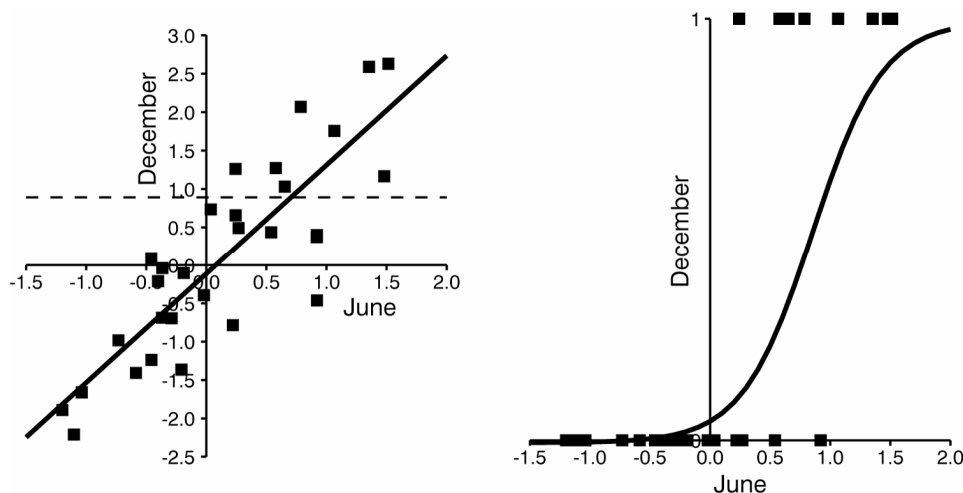


Figure 7.4 – Example of (a) a linear regression model and (b) a generalised linear regression model. June values of the Niño3.4 index are used to predict December 1971–2000 values. The dashed horizontal line represents the upper quartile of December values of the index.

7.4.3.b Classification procedures

Classification procedures have been used in seasonal climate forecasting more extensively than generalised linear models. As with generalised linear models, the observations are assigned to one of two or more categories, and then probabilities are calculated that a new observation will be within each of the categories given new values of the predictors. An important distinction, however, is that categories are nominal in classification procedures, so that if there are three or more, the procedures do not know, for example, that they are ordered as below-normal, near-normal, and above-normal. In most cases of seasonal climate forecasting the fact that the categories are nominal in classification procedures is likely to be a disadvantage because relationships between predictors and predictands are most often likely to be monotonic.

Discriminant analysis is the most widely used classification procedure in seasonal climate forecasting. The values of the predictand are assigned to one of the categories, and the mean values of the predictors are then calculated for each category separately. If the predictors have good discriminatory power then the differences in the means of the predictors between the various categories will be large. For example, if seasonal rainfall is strongly influenced by the ENSO phenomenon, then the difference in the average value of the Niño3.4 index when rainfall is above-normal compared to when rainfall is below-normal will be large. Given the covariances of the predictors in each category the probability that a new observation will be in each category can be calculated from the new values of the predictors, and from knowledge about the prior probabilities of each category. Mathematically, it is simpler to assume that the covariances are the same for each category, and a linear classification can be defined to identify the most likely category. If this assumption of equal covariance is dropped, the classification function becomes quadratic. The quadratic function only performs noticeably better than the linear function when the differences in covariance are marked.

Canonical variate analysis has had limited application in seasonal climate forecasting, but it has been used in predicting the phase of the ENSO phenomenon. The technique is similar to discriminant analysis, but has some similarities to canonical correlation analysis as well. Just as canonical correlation analysis identifies optimal linear combinations of the predictors to maximise correlations with linear combinations of the predictands, canonical variate analysis seeks optimal linear combinations of the predictors, but in this case to maximise the discrimination between the categories. The discrimination is defined by the ratio of between-group to total variance.

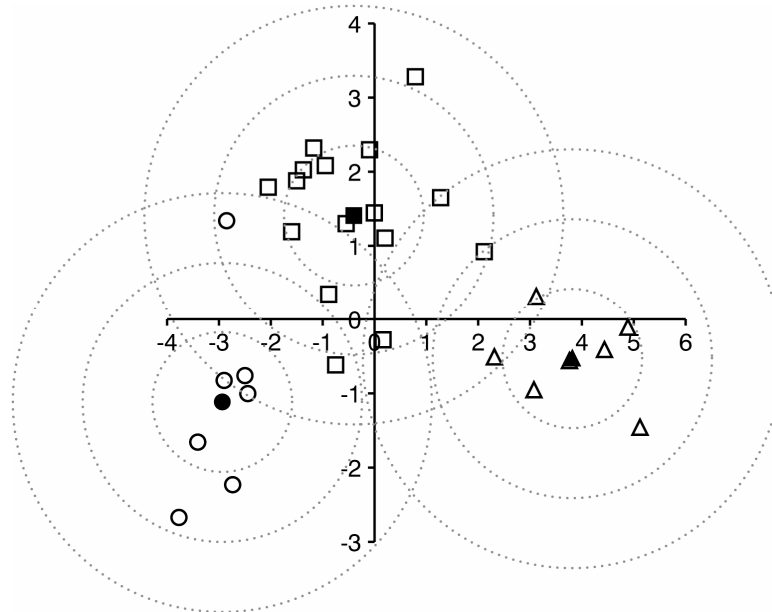


Figure 7.5 – Example of a canonical variate analysis model. The x -axis represents the first canonical variate of monthly Niño3.4 indices from January–November, and the y -axis the second. The hollow symbols represent observed scores on the canonical variates for 1971–2000, and the solid symbols the corresponding mean values. The circles represent years in which the December Niño3.4 index was below the lower quartile, the triangles years in which the index was above the upper quartile, and squares years in which the index was within the inter-quartile range. The large dashed circles represent distances of one standard deviation.

An example is provided in Figure 7.5, where canonical variates are computed using monthly Niño3.4 indices from January–November to predict the ENSO phase for the following December. Three phases are defined based on the outer quartiles of the December value of the index, and are represented by the different symbols: the open circles represent years in which the December Niño3.4 index was below the lower quartile (i.e. La Niña events), the open triangles years in which the index was above the upper quartile (i.e. El Niño events), and the open squares years in which the index was within the inter-quartile range (i.e. neutral events). The x -axis represents the first canonical variate (a linear combination of the Niño3.4 indices for January–November), which maximises the distances between the mean values of canonical variate scores for the three categories, as represented by the solid symbols. This canonical variate therefore maximises the distances along the x -axis between the three solid symbols. The first canonical variate successfully distinguishes the three categories, but is most effective in identifying the El Niño events (represented by the triangles).

The second canonical variate maximises the distances between the categories along the y -axis, and helps to distinguish the La Niña events (circles) from the neutral events (squares). The dashed circles indicate distances in multiples of one standard deviation from the category means, (assuming that the variances in all three categories are equal), and can be used to visualise in which category a new observation is most likely to occur.

Classification procedures address a number of the problems listed in section 7.4.1. Because the predictands are categorised in both discriminant analysis and canonical variate analysis, no assumptions are made about their distribution. However, it is assumed that the predictors are normally distributed, and linear discriminant analysis is sensitive to violations of this assumption. Quadratic analysis is more robust, except when the data are highly skewed. As with the forms of generalised linear models discussed in section 7.4.3.a, multiplicity and multicollinearity remain as problems, but can be addressed in similar ways to that for linear regression, e.g. by using principal components as predictors.

7.4.3.c *Analogue procedures*

Analogue procedures have some similarity to classification procedures, but are listed separately because of a number of important differences from discriminant analysis and canonical variate analysis, and because of a wide flexibility in how the analogues can be used to make a prediction. The essential step is to identify years from the historical records in which the states of the predictors were similar to the states for the current forecast. Some index of similarity (or of dissimilarity) is used to calculate how closely current conditions resemble previously observed conditions. A frequently used measure of similarity is the Mahalanobis distance, which is similar to the squared distance, but which compensates for correlations between the predictors.

The distinction between this step of identifying similar years and classification is that the similarity of individual years, rather than of the mean of a predefined category of years, is investigated. However, in some of the simpler analogue procedures, often, but not exclusively, used when there is only one predictor, the predictor(s) is (are) classified into one of a set of predefined classes, and other years within this category are treated as analogues. A widely used example of this classification step in an analogue procedure is the Southern Oscillation phase system, in which the current state and recent evolution of the Southern Oscillation Index are classified into one of the five categories rapidly falling, rapidly rising, consistently positive, consistently negative, and consistently near-zero.

Once analogue years have been identified, a forecast is constructed using the observed values for these selected years. The forecast can be

constructed in a number of ways, the simplest of which is to use the mean value, although normally the variability within the analogue years would also be considered to provide some indication of the uncertainty in the forecast. If the forecast sample is sufficiently large, the probability that the predictand will exceed a threshold value could be obtained by counting the proportion of times it was exceeded in the analogue sample (although errors in calculating this proportion are likely to be large). A more reliable approach would be to fit an appropriate distribution to the analogue and to derive a forecast from this fitted distribution. The problem is essentially identical to that of constructing a forecast from an ensemble of GCM predictions. Each analogue year can be treated as an ensemble member. Procedures for obtaining a forecast from an ensemble are discussed in Chapter 8 (section 8.5.2).

A special case of an analogue procedure is the constructed analogue, which combines *all* previous cases. The procedure is a form of ridge regression, which is discussed in further detail in section 7.4.2.