6

Model Adequacy: Determining a Network's Future Performance

In this chapter we present various metrics in order to assess a trained network. We are interested in measuring the predictive ability of a wavelet network in the context of a particular application. The evaluation of the model usually includes two clearly distinct, although related stages.

In the first stage, various metrics that quantify the accuracy of the predictions or the classifications made by the model are used and the model is evaluated based on these metrics. The term *accuracy* is a quantification of the "proximity" between the outputs of the wavelet network and the target values desired. The measurements of the precision are related to the error function that is minimized (or in some cases, the profit function that is maximized) during model specification of the wavelet network model. When an estimate of the model is done by minimizing the squared error function, the simplest example of such a measurement of accuracy is the mean squared error (MSE). The most common error criteria are the MSE, the root MSE, the normalized MSE, the sum of squared errors, the maximum absolute error, the mean absolute error, the (symmetric) mean absolute percentage error, and Theil's U index.

In addition, useful and immediate information can be provided by visual examination of a scatter plot of the network forecasts and the target values. Moreover, the statistical hypothesis testing of the values of the intercept and the slope of the regression between the wavelet network outputs and the target values can also provide useful information. Indicators such as the prediction of change in direction, independent prediction of change in direction, and the position of sign evaluate the ability of a wavelet network to predict changes in sign or direction of the values. In

Wavelet Neural Networks: With Applications in Financial Engineering, Chaos, and Classification, First Edition. Antonios K. Alexandridis and Achilleas D. Zapranis.

classification applications, metrics such as the classification rates, relative entropy, and the Kolmogorov–Smirnov statistic are used. The metrics above provide useful information about the predictive capabilities of the wavelet network; however, they are not sufficient for a complete evaluation of the adequacy of our model.

The second step is to assess the behavior and the performance of the wavelet network model under conditions as close as possible to the actual operating conditions of the application. The greater accuracy of the network model does not necessarily mean that it will be applied more successfully. For example, in a time-series application for forecasting returns, it is possible for a model that is characterized by low MSE to create large losses because of only a few failed predictions that are accompanied by high costs (Yao et al., 1999).

It is important, therefore, that the performance of the model is evaluated in the context of decision making that it supports. Especially in the case of time-series forecasting applications, the performance and the evaluation of the model should be based on benchmark trading strategies. It is also important to evaluate the behavior of the model throughout the range of the actual possible scenarios. For example, a trader should know how a predictive model behaves during an upward or downward trend in the market, sideways movements, if it is able to predict turning points, and so on.

The wavelet neural model is evaluated using validation samples. The patterns used in the validation samples were not used during the training of the wavelet network and represent all possible scenarios that can arise in reality. For example, if the application concerns prediction of the performance of a stock index, it would not be correct if the validation sample corresponds solely to a period with a strong upward trend since the evaluation will be restricted to the specific circumstances. It is possible to use more than one validation sample, each corresponding to different conditions, to cover the full range of possible scenarios.

The full understanding of the behavior of the model under different conditions is a prerequisite for the creation of a successful decision support system or simply a decision-making system (e.g., automated trading systems). The sensitivity analysis of the model will help us understand the dynamics and the nature of the process that the wavelet network has learned during the training phase. An optical analysis of the relationship between the explanatory and the dependent variable is recommended, by constructing two- or three-dimensional plots that connect the value of the dependent variable with the value of one or two explanatory variables, respectively. Using the model adequacy process and the appealing properties of wavelet networks, the wavelet neural model ceases to be a "black box" but, rather, it constitutes a reliable framework where we can base our decisions depending on the application. Almost always, a pilot phase follows where the final evaluation of our model is performed under the actual conditions.

TESTING THE RESIDUALS

For a "correctly specified" wavelet neural network model, the nonparametric residuals

$$y_i - g(\mathbf{x}; \hat{\mathbf{w}}_n) = e_i \tag{6.1}$$

are such that $e_i \cong \varepsilon_i = y_i - \varphi(x_i)$. The residuals $\{e_i\}$ can be used to perform meaningful diagnostic tests about the initial assumptions regarding error term ε of the theoretical underlying model $y_i = \varphi(x_i) + \varepsilon_i$. However, because of the nonparametric nature of wavelet neural networks, satisfying those tests is a "necessary but not sufficient" condition for model adequacy. As in the case of the methodology proposed by Box and Jenkins (1970) for ARMA models, the stage of diagnostic checking should be integrated in the process of specifying a model, but it cannot replace it.

The graphical representation and visual inspection of the residuals can reveal extreme values, autocorrelations, or periodicities. To test for autocorrelation, the true nature of the serially correlated error process must be known. However, often this is not the case. As a result, the methods proposed include the fitting of stationary time series to ordinary least squares residuals. An example is the Durbin–Watson test, which has been developed for linear regression models to test the hypothesis that $\varphi = 0$ in the error model $e_i = \varphi e_{i-1} + a_i$. For linear models it has been shown that the autocorrelations estimated from the OLS residuals converge to the true autocorrelations. This observation can lead to tests such as the Box–Pierce and Ljung–Box, which are used for adequacy testing of ARMA models. These tests will still hold asymptotically for nonlinear models such as wavelet neural networks, although larger samples will be needed for finite sample validity (Zapranis and Refenes, 1999). It should be noted here that the relevant theory behind these tests refers to an observed time series. Here we use $\{e_i\}$ and we assume that it is sufficiently close to $\{\varepsilon_i\}$. In the following paragraphs we refer to various diagnostic tests, including a test for the adequacy of the model in the sense of heteroscedasticity presence in the error term. More sophisticated tests to identify the source of the specification error also exist. One typical example is Ramsey's RESET test (Ramsey, 1969), which is based on the residuals of estimated linear regression models, and it is used to identify specification errors due to omitted variables or incorrect functional form. A similar test for neural networks was proposed by White (1989).

Testing for Serial Correlation in the Residuals

Correlogram The autocorrelation function (ACF) can be used to determine whether there is any pattern in the residuals, and the absence of such a pattern may reveal that the particular sample of residuals is random. The sample autocorrelation function for lag k, denoted by $\hat{r}_k$, is defined as follows:

$$\hat{r}_k = \frac{\hat{\gamma}_k}{\hat{\gamma}_0} \tag{6.2}$$

where γ_k is the sample covariance (or autocovariance) at lag k and γ_0 is the sample variance (i.e., the covariance at lag 0), which are given by the following two equations:

$$\hat{\gamma}_k = \frac{\sum (e_i - \bar{e})(e_{i+k} - \bar{e})}{n} \tag{6.3}$$

$$\hat{\gamma}_0 = \frac{\sum (e_i - \bar{e})^2}{n} \tag{6.4}$$

where $\bar{e}$ is the mean value of the residuals and n is the sample size. A plot of $\hat{r}_k$ against k is known as the sample correlogram. It can be shown that if a time series is purely random, the sample autocorrelation coefficients are asymptotically normally distributed with zero mean and variance $1/n$. Thus, it can be concluded that the residuals are random if the autocorrelations calculated are within the limits

$$-z_a \frac{1}{\sqrt{n}} \leq \hat{r}_k \leq +z_a \frac{1}{\sqrt{n}} \tag{6.5}$$

with z_a being the $100(1-a)$ percentile point of the standard normal distribution.

Box–Pierce* Q-*Statistic To test the joint hypothesis that all the autocorrelation coefficients are simultaneously equal to zero (i.e., to test the null hypothesis H_0: $r_1 = r_2 = \cdots = r_m = 0$ against the alternative that not all autocorrelation coefficients are zero), one can use the Box–Pierce test (Makridakis et al., 2008). The Box–Pierce test requires computation of the "portmanteau" Q-statistic:

$$Q = n \sum_{k=1}^{m} \hat{r}_k^2 \tag{6.6}$$

where m is the lag length. Under H_0, Q is distributed asymptotically as χ_m^2.

Ljung–Box LB-Statistic A variant of the Box–Pierce Q-statistic is the Ljung–Box LB-statistic, defined as

$$\text{LB} = n(n+2) \sum_{k=1}^{m} \frac{\hat{r}_k^2}{n-k} \tag{6.7}$$

which under the null hypothesis is also distributed as χ_m^2. The Ljung–Box test has better properties than the Box–Pierce test. Typically, the Q-statistic is evaluated for several choices of m. Under H_0, for large n,

$$E[Q(m_1) - Q(m_2)] = m_2 - m_1 \tag{6.8}$$

so that different sections of the correlogram can be checked for departures from H_0. Furthermore, acceptance of H_0 requires one to check the individual autocorrelation coefficients to see whether a large number of them are close to zero and mask the presence of a highly significant individual.

Durbin–Watson Statistic The Durbin–Watson test was derived for linear regression models to test the hypothesis that $\varphi = 0$ in the error model $e_i = \varphi e_{i-1} + a_i$. The test statistic is

$$d = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=2}^{n} e_i^2} \approx 2(1 - \hat{r}_1) \tag{6.9}$$

where $\hat{r}_1$ is the sample autocorrelation at lag $k = 1$. Durbin and Watson (1951) calculated bound d_L and d_U for d, which depend on the number of linear regressors. The null hypothesis $H_0: \varphi = 0$ versus the alternative $H_1: \varphi > 0$ is not rejected if $d > d_U$, and it is rejected if $d < d_L$. The test is inconclusive if $d_L < d < d_U$. To test versus $H_1: \varphi < 0$, we replace d with $4 - d$. The test was described analytically by Judge et al. (1982). The asymptotic equivalence (locally) of a nonlinear regression [e.g., the wavelet network model $g(\mathbf{x}; \hat{\mathbf{w}}_n)$] and its linear approximation around the true parameter vector $\mathbf{w}_0$, as given by the first-order Taylor approximation

$$g(\mathbf{x}; \mathbf{w}_0) + \frac{\partial(\mathbf{x}; \mathbf{w}_0)}{\partial \hat{\mathbf{w}}_n^T}(\hat{\mathbf{w}}_n - \mathbf{w}_0) \tag{6.10}$$

suggests that the test will be approximately valid. However, the applicability of the Durbin–Watson test for nonlinear models has not been established rigorously (Zapranis and Refenes, 1999).

EVALUATION CRITERIA FOR THE PREDICTION ABILITY OF THE WAVELET NETWORK

To evaluate the ability of the wavelet network to predict the level of values, prices, returns, or trends, various metrics are used. Generally, we can distinguish between two groups of measurements: (1) measurements regarding the accuracy of the prediction of our model in isolation or compared against another reference model, and (2) measurements regarding the predictability of the changes in direction of the values. Also, very often, visual inspection of the scatter plot between the wavelet network predictions against target values is used, as it provides immediate information. Moreover, in this section we present and analyze the predictive power of the wavelet network based on statistical hypothesis testing of the values of the parameters of the linear regression between the outputs of the wavelet network and the target values.

Measuring the Accuracy of the Predictions

The usual metric that is used for quantification of the accuracy of the predictions of the network is the mean squared error (MSE). To find the MSE, first the difference between the network predictions and the real target values is estimated. This difference is squared and then the average is taken. The MSE is given by

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{6.11}$$

where y_i are the target values, $\hat{y}_i$ are the wavelet networks prediction, and n is the number of observations.

Another measure that is often used is the root mean squared error (RMSE). The RMSE is simply the squared root of the MSE. Since MSE is considered the variance

of an unbiased estimator, in the same sense, the RMSE is the standard deviation. The RMSE is given by

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} = \sqrt{\text{MSE}} \tag{6.12}$$

An extension of the previous metric is the normalized mean squared error (NMSE), which is given by

$$\text{NMSE} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{6.13}$$

where $\bar{y}$ is the average value of the target values:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{6.14}$$

On the other hand, the mean absolute error (MAE) measures the accuracy in absolute terms:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{6.15}$$

A metric similar to the MAE is the median absolute error (Md.AE). The median is the numerical value that separates the higher half of the errors from the lower half. Hence, to estimate the Md.AE, we first arrange all the absolute errors from the lowest value to the highest, and then we pick the middle one. If there is an even number of observations, there is no single middle value; in this case the Md.AE is defined to be the mean of the two middle values.

Sometimes the sum of squared errors (SSE) is used. The SSE is given by

$$\text{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{6.16}$$

In some cases it is useful to estimate the maximum absolute error (MaxAE):

$$\text{MaxAE} = \max\left|y_i - \hat{y}_i\right| \tag{6.17}$$

Similarly, the mean absolute percentage error (MAPE) measures the accuracy of the network prediction in absolute and percentage terms:

$$\text{MAPE} = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{6.18}$$

The difference between the actual value and the prediction is again divided by the actual value. However, the concept of MAPE has a major drawback in practical applications. If there are zero values in the data sample, there will be a division by zero. In the case of a perfect fit, MAPE is zero.

The symmetrical mean absolute percentage error (SMAPE) is an extension of the MAPE. The SMAPE is given by

$$\text{SMAPE} = \frac{100}{n} \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{(y_i + \hat{y}_i)/2} \tag{6.19}$$

This formula provides a result between 0 and 200%. However, a percentage error between 0 and 100% is much easier to interpret. Hence, the following formula is used in practice:

$$\text{SMAPE} = \frac{100}{n} \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{y_i + \hat{y}_i} \tag{6.20}$$

Often, we want to compare the performance between two models. To compare the predictive power of a wavelet network and a benchmark, Theil's U Index is used:

$$U = \sqrt{\sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{y_i - f_i} \right)^2} \tag{6.21}$$

where f_i are the predictions of the benchmark model. If $U = 1$, the prediction power of the wavelet network is equal to the predictive power of the benchmark model. If $U > 1$, the performance of the networks is worse than the benchmark model, while if $U < 1$, the network's performance is better. The U index is generally used in forecasting econometric time series, and the benchmark model is the simple random walk.

Scatter Plots

The visual examination of a scatter plot between the forecasts of the wavelet network and the target values can provide direct information about the predictive ability of the network. In a scatter plot, the horizontal axis represents the predictions of the wavelet network, $\hat{y}_i$, and the vertical axis represents the target values, y_i. Each point on the graph corresponds to a pair of values $(y_i, \hat{y}_i)$. Clearly, it is desirable that the points are to be distributed near and around the straight line $y_i = \hat{y}_i$ which has a slope of 45° and passes through the origin of the axes. In Figures 6.1a and 6.2 we see this case exactly. On the other hand, Figure 6.1b shows a model that generates predictions rather randomly. An intermediate situation between the two corresponds to an intermediate predictor. Other possible cases are presented in Figure 6.1c and d. In the first case we have a linear relationship but with a slope different from 45°; in the second case the network has no predictive power and almost always returns the same output value.

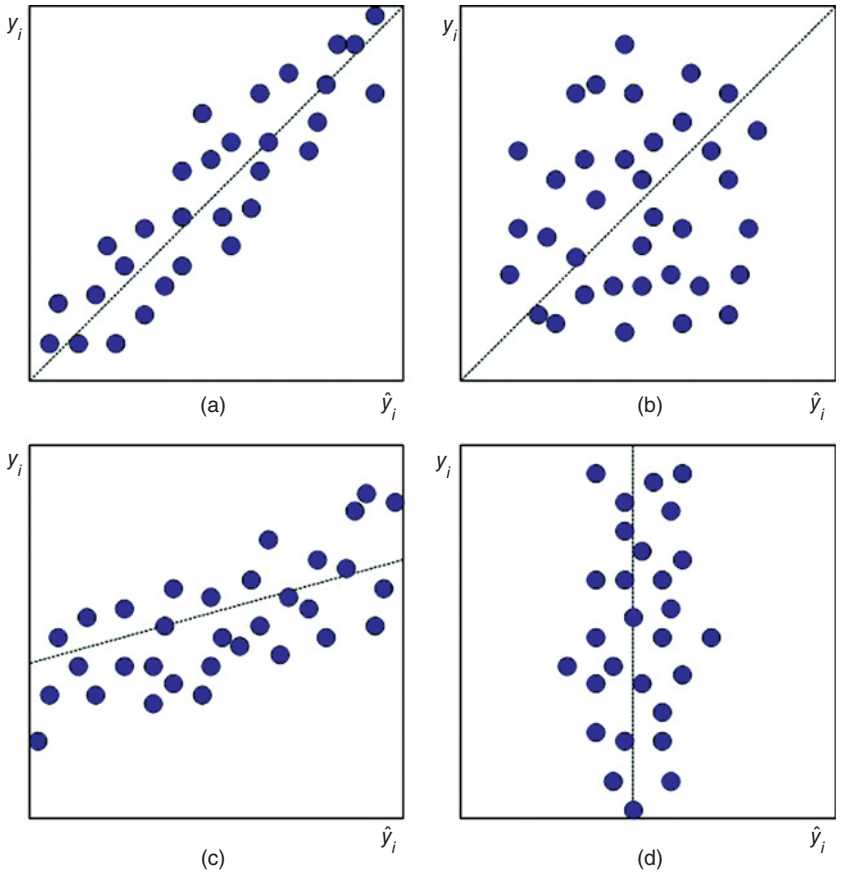

Figure 6.1 *Various scatter plots between the wavelet network's predictions and the target values.*

Linear Regression Between Forecasts and Targets

Although scatter plots provide an easy and direct (visual) assessment of the predictive power of the wavelet network, which do not quantify this ability. A very useful way to quantify the forecasting ability of a wavelet network, which also allows comparison of different models, is the linear regression between the forecasts of the wavelet network and the target values: in other words, estimation of the following model:

$$y_i = \beta_0 + \beta_1 \hat{y}_i \tag{6.22}$$

The estimated parameters that arise from the OLS of the parameters β_0 and β_1 are denoted by b_0 and b_1, respectively. The residuals of the regression are given by $e_i = y_1 - (b_0 + b_1 \hat{y}_i)$. Hence, the estimated model is

$$y_1 = b_0 + b_1 \hat{y}_i + e_i \tag{6.23}$$

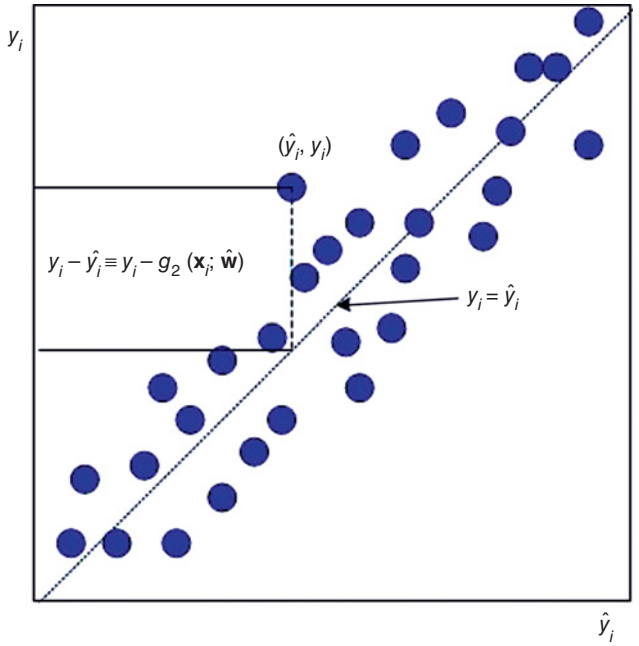

Figure 6.2 *Scatter plot between the wavelet network's predictions and the target values where the line* $y_i = \hat{y}_i$ *passes through the origin and has a slope of* 45°.

To estimate the *t*-statistics and *p*-values of the parameters b_0 and b_1, a normal distribution is assumed.

If the slope of the regression b_1 is equal to zero, there is no linear relationship between the targets y_i and the predicted values $\hat{y}_i$. There are two cases where the slope of the regression can be zero:

- When the predicted value, $\hat{y}_i$, is constant for every target value.
- When the predicted values, $\hat{y}_i$, and the targets, y_i, are completely uncorrelated.

In any other case there is a form of linear relationship. The coefficient of determination R^2 indicates how well data points fit a line or curve. The values of R^2 are between 0 and 1 and given by the following relationship:

$$R^2 = 1 - \frac{\mathrm{SSE}_{\mathrm{TP}}}{\mathrm{SST}_{\mathrm{TP}}} \tag{6.24}$$

where $\mathrm{SSE}_{\mathrm{TP}}$ is the sum of the squared errors given by

$$\mathrm{SSE}_{\mathrm{TP}} = \sum_{i=1}^{n} [y_i - (b_0 + b_1 \hat{y}_i)]^2 = \sum_{i=1}^{n} e_i^2 \tag{6.25}$$

and SST_{TP} is the total sum of squares:

$$\text{SST}_{\text{TP}} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{6.26}$$

We use the subscript TP to refer to the regression between the target values and the values predicted.

The F-statistic informs us if there is a linear relationship between the target values and the values predicted:

$$F = \frac{\text{SST}_{\text{TP}} - \text{SSE}_{\text{TP}}}{\text{SSE}_{\text{TP}}/(n-2)} \tag{6.27}$$

The same information can be obtained from the t-statistic of the slope of the regression b_1. More precisely, the t-statistic is used for the following hypothesis test regarding the existence of a linear relationship between the target values, y_i, and the values predicted, $\hat{y}_i$:

$$\begin{aligned} H_0&: b_1 = 0 \\ H_1&: b_1 \neq 0 \end{aligned} \tag{6.28}$$

If the slope of the linear regression is zero, there is no linear relationship between the targets y_i and the predictions $\hat{y}_i$. If it is not equal to zero, the slope will be either positive or negative, and as a result, a linear relationship will exist.

Assuming that the residuals e_i follow a normal distribution, the t-statistic follows the t-Student distribution with $\upsilon = n - 2$ degrees of freedom and is given by

$$t_{n-2} = \frac{b_1}{\text{s.e.}(b_1)} \tag{6.29}$$

where s.e. (b_1) is the standard error of the slope b_1, which is given by

$$\text{s.e.}(b_1) = \frac{s}{\sqrt{\sum_{i=1}^{n} (\hat{y}_i - \hat{\bar{y}}_i)}} \tag{6.30}$$

where s is the standard error of the residuals:

$$s = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}} = \sqrt{\frac{\text{SSE}_{\text{TP}}}{n-2}} \tag{6.31}$$

and

$$\hat{\bar{y}} = \frac{1}{n} \sum_{i=1}^{n} \hat{y}_i \tag{6.32}$$

For a large sample and a significance level $\alpha = 0.05$, the critical values of the t-Student distribution are ± 1.96. If the t-statistic is larger than 1.96 or smaller than -1.96, the hypothesis H_0 is rejected. In other words, b_1 is statistically significantly different from zero; hence, a linear relationship exists between the targets y_i and the predictions $\hat{y}_i$. Alternatively, the p-values can be calculated. The p-values report the minimum significance level at which the null hypothesis can be rejected.

The previous test is not sufficient for our analysis. Even if a linear relationship exists between the targets y_i and the predictions $\hat{y}_i$, it is possible that the imaginary line formed by the pairs $(\hat{y}_i, y_i)$ does not pass through the origin of the axes or have a slope different from 45°, or both.

When the imaginary straight line does not pass through the origin, the constant term of the regression b_0 is different from zero. Therefore, in this case the null hypothesis H_0 that the constant term is zero should be tested against the alternative H_1 that it is different from zero:

$$\begin{aligned} H_0&: b_0 = 0 \\ H_1&: b_0 \neq 0 \end{aligned} \tag{6.33}$$

The test of the hypothesis of a zero value of b_0 is based on the t-ratio

$$t_{n-2} = \frac{b_0}{\text{s.e.}(b_0)} \tag{6.34}$$

where s.e.(b_0) is the standard error of the constant b_0, which is given by

$$\text{s.e.}(b_1) = s\sqrt{\frac{1}{n} + \frac{\hat{\bar{y}}_i^{\,2}}{\sum_{i=1}^{n}(\hat{y}_i - \hat{\bar{y}}_i)^2}} \tag{6.35}$$

The t-ratio given by equation (6.34) is compared against the critical value of the t-Student distribution with $\upsilon = n - 2$ degrees of freedom.

Finally, we must test whether or not the slope of the imaginary straight line formed by the pairs $(\hat{y}_i, y_i)$ is different from 45°. In other words, we want to test the null hypothesis H_0, that the slope of the regression $b_1 = 1$, against H_1, that $b_1 \neq 1$:

$$\begin{aligned} H_0&: b_1 = 1 \\ H_1&: b_1 \neq 1 \end{aligned} \tag{6.36}$$

The t-ratio given by (6.29) is a special version of the statistic

$$t_{n-2} = \frac{b_1 - \beta_{1,0}}{\text{s.e.}(b_1)} \tag{6.37}$$

where $\beta_{1,0}$ is the value of the slope b_1 under the null hypothesis. Hence, for $b_1 = 1$ we have that

$$t_{n-2} = \frac{b_1 - 1}{\text{s.e.}(b_1)} \tag{6.38}$$

It should be noted that if autocorrelation exists in the residuals of the regression e_i, the standard errors of the intercept b_0 and the slope b_1 of the regression will be very small. As a result, it will be difficult to accept the null hypotheses H_0: $b_0 = 0$ and H_0: $b_1 = 1$. For this reason, the statistical Durbin–Watson (DW), which is used for the diagnosis of autocorrelation in the residues of regression, is also reported.

Summing up, for a properly identified, nonbiased wavelet neural network, a linear relationship between the targets and the predictions of the network should exist, the intercept b_0 should be equal to zero, and the slope of the regression b_1 should be equal to 1. It should be clarified that this is a necessary but not a sufficient condition for a proper model identification framework. For example, an overparameterized wavelet neural network model that "learned" the noise that exists in the training data, but not the underlying function that generated the observations, will satisfy the foregoing conditions for the training sample but will be biased with reduced generalization ability to new data.

The generating process of the observations is given by the relationship

$$y_i = \varphi(\mathbf{x}_i) + \varepsilon_i. \tag{6.39}$$

where the error ε_i is distributed identically and independently with mean zero and variance σ_ε^2. We have that

$$\begin{aligned} &\text{var}\left[(y_i - \hat{y}_i)^2\right] = \text{var}\left[(\varphi(\mathbf{x}_i) - \hat{y}_i)^2\right] + \text{var}\left[\varepsilon_i^2\right] \\ &\Leftrightarrow \sigma_p^2(\mathbf{x}) = \sigma_m^2(\mathbf{x}) + \sigma_e^2 \end{aligned} \tag{6.40}$$

where σ_p^2 is the prediction variance and σ_m^2 is the model variance.

As can be seen from equation (6.40) the size of the dispersion of the pairs $(\hat{y}_i, y_i)$ around the imaginary straight $y_i = \hat{y}_i$ depends on two factors: (1) how well the wavelet neural model was identified [i.e., how close the predictions of the network $\hat{y}_i$ are to the unknown underlying function $\varphi(\mathbf{x}_i)$] and (2) the variance of the error term σ_ε^2. It follows that the pairs $(\hat{y}_i, y_i)$ will lie on the imaginary straight line $y_i = \hat{y}_i$ if the model has been specified properly and there is no error term.

Measuring the Ability to Predict the Change in Direction

While the indicators discussed previously examine the accuracy of the prediction, in some cases we are also interested in predicting the changes in direction of the dependent value independent of their range. These metrics are often used in the analysis and forecasting of financial time series and are expressed as percentages.

The first indicator is the prediction of change in direction (POCID):

$$\text{POCID} = \frac{100}{n} \sum_{i=1}^{n} d_i \tag{6.41}$$

where

$$d_i = \begin{cases} 1 & (y_i - y_{i-1})(\hat{y}_i - y_{i-1}) > 0 \\ 0 & (y_i - y_{i-1})(\hat{y}_i - y_{i-1}) \leq 0 \end{cases} \tag{6.42}$$

An extension of the POCID is the independent prediction of change in direction (IPOCID), given by

$$\text{IPOCID} = \frac{100}{n} \sum_{i=1}^{n} d_i \tag{6.43}$$

where

$$d_i = \begin{cases} 1 & (y_i - y_{i-1})(\hat{y}_i - \hat{y}_{i-1}) > 0 \\ 0 & (y_i - y_{i-1})(\hat{y}_i - \hat{y}_{i-1}) \leq 0 \end{cases} \tag{6.44}$$

Finally, an indicator often used in forecasting financial returns is the prediction of sign (POS), given by

$$\text{POS} = \frac{100}{n} \sum_{i=1}^{n} d_i \tag{6.45}$$

where

$$d_i = \begin{cases} 1 & y_i \cdot \hat{y}_i > 0 \\ 0 & y_i \cdot \hat{y}_i \leq 0 \end{cases} \tag{6.46}$$

TWO SIMULATED CASES

In this section we demonstrate how the previous metrics can be used in two simulated cases to assess the model adequacy of a wavelet network. The first case is the sinusoid with a decreasing variance in the noise, and the second case is a summation of two sinusoids with Cauchy noise. Both cases were discussed in earlier chapters.

Case 1: Sinusoid and Noise with Decreasing Variance

As discussed in Chapters 3 and 4, a wavelet network with 2 hidden units was used. The initialization method was presented in Chapter 3 and various methods for selecting the optimal architecture of the wavelet network were presented in Chapter 4. In this

TABLE 6.1 Case 1: Residuals Testing[a]

	Parameter	p-Values
n/p Ratio	125.1250	
Mean	0.0001	
Median	0.0045	
S. dev.	0.1771	
DW	1.9594	0.5199
LB Q-stat.	42.5310	0.0024
JB stat.	74.26664	0.0010
KS stat.	10.5471	0.0000
R^2	0.71044	
$\bar{R}^2$	0.70840	

[a]S. dev., standard deviation; DW, Durbin–Watson; LB, Ljung–Box; KS, Kolmogorov–Smirnov.

section we examine how well the wavelet network was trained on the data: in other words, the ability of the network to learn the data and then forecast the future values of the underlying function.

Testing the Residuals First, the residuals of the trained networks will be examined. In Table 6.1 the various descriptive statistics as well as various tests on the residuals are presented. The mean of the residuals is close to zero with a standard deviation of 0.17. Both the Durbin–Watson and the Ljung–Box tests reject the hypothesis of uncorrelated residuals. Similarly, the Komlogorov–Smirnov and Jarque–Berra statistics indicate the absence of normality in the residuals, which is logical taking into account the generating process of the residuals. Finally, both the R^2 and the $\bar{R}^2$ show a good fit of the network to the data.

Error Criteria Various error criteria are presented in Table 6.2. These criteria will help us assess the fit of the wavelet network to the data: in other words, how well the wavelet network learned the data. All criteria are very small, with MSE and NMSE 0.0313 and 0.286, respectively. On the other hand, the MAPE and SMAPE are 119.39% and 44.74%, respectively. Note that there is a large presence of noise to the data. As a result, the error criteria will increase. The wavelet network learned the underlying function successfully as presented in earlier chapters without being affected by the noise.

Scatter Plot Figure 6.3 is a scatter plot of the real values versus the values forecasted

TABLE 6.2 Case 1: Error Criteria[a]

Md.AE	MAE	MaxAE	SSE	RMSE	NMSE	MSE	MAPE	SMAPE
0.0866	0.1276	0.6813	31.3626	0.1770	0.2896	0.0313	119.39%	44.74%

[a]Md.AE, median absolute error; MAE, mean absolute error; MaxAE, maximum absolute error; SSE, sum of squared errors; RMSE, root mean square error; NMSE, normalized mean squared error; MSE, mean square error; MAPE, mean absolute percentage error; SPAME, symmetric mean absolute percentage error.

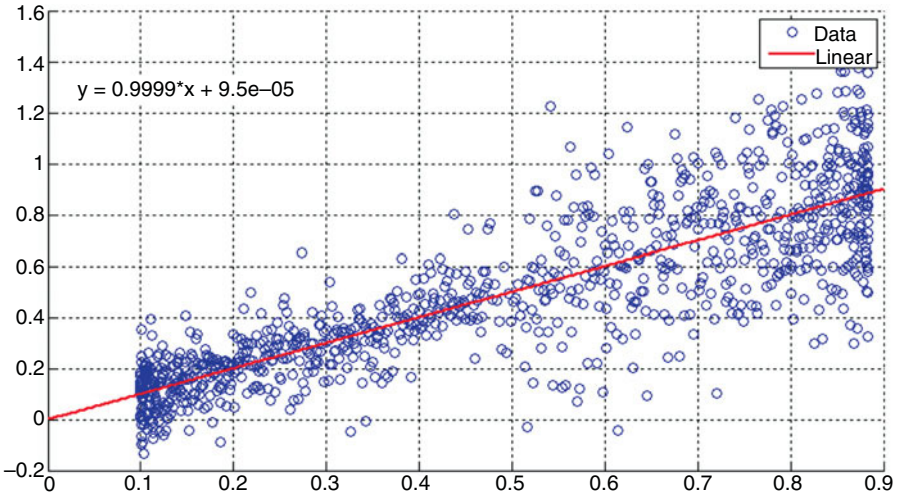

Figure 6.3 *Scatter plot for the first simulated case.*

for the wavelet network. As can be seen, the forecasts and the targets are in line, but there is a large dispersion, caused by the large noise term.

Regression Between the Forecasts and the Target Values The estimated parameters of the regression between the forecasted values and the target values are presented in Table 6.3. Close inspection reveals that parameter b_0 is not statistically different from zero, while the parameter b_1 is not statistically different from 1. Moreover, the linear regression is statistically significant, according to the F-statistic.

Changes in Direction The POCID, IPOCID, and POS criteria are presented in Table 6.4. In this application the aim is to learn the underlying process that is masked by a large noise part with increasing variance. As a result, we expect the POCID and IPOCID to be relatively low. More precisely, the POCID and IPCOID are 75.4% and 50.9%, respectively, while the POS is 97.5%.

TABLE 6.3 Case 1: Regression Statistics for the First Simulated Case[a]

	Parameter	p-Values	S.E.	T-Stat.
b_0	0.0001	0.9934	0.0114	0.0083
b_1	1.0000	0.0000	0.0202	49.5080
$b_1 = 1$ test	1.0000	0.9996	0.0202	−0.0004
R^2	0.7104			
F	2451.1000	0.0000		
DW	1.9594	0.49959		

[a]S.E., squared errors; DW, Durbin–Watson.

TABLE 6.4 Case 1: Change-in-Direction Metrics[a]

POCID	IPOCID	POS
75.4%	50.9%	97.5%

[a]POCID, prediction of change in direction; IPOCID, independent prediction of change in direction; POS, prediction of sign.

TABLE 6.5 Case 1: Out-of-Sample Residual Testing[a]

	Parameter	p-Values
n/p Ratio	37.5000	
Mean	0.0031	
Median	−0.0089	
S. dev.	0.2211	
DW	1.8484	0.1876
LB Q-stat.	36.7980	0.0124
JB stat.	5.2543	0.0617
KS stat.	5.3766	0.0000
R^2	0.6578	
$\bar{R}^2$	0.6496	

[a]S. dev., standard deviation; DW, Durbin–Watson; LB, Ljung–Box; KS, Kolmogorov–Smirnov.

Out-of-Sample Forecasts In this section the out-of-sample performance of the wavelet network is evaluated. The various descriptive statistics as well as various tests on the out-of-sample residuals are presented in Table 6.5. The mean of the residuals is close to zero with a standard deviation of 0.22. Both the Durbin–Watson and Ljung–Box tests reject the hypothesis of uncorrelated residuals. Similarly, the Komlogorov–Smirnov test indicates the absence of normality in the residuals. On the other hand, the normality hypothesis is accepted at a 5% confidence level using the Jarque–Berra statistic. Hence, comparing Tables 6.1 and 6.5, we can conclude that there are similar circumstances that generated the training and the validation sample. Finally, both R^2 and $\bar{R}^2$ show the good prediction ability of the network, considering the large amount of noise.

Next, the accuracy of the prediction is evaluated. In Table 6.6 various error criteria are presented. The values of the metrics are similar to those presented in Table 6.2. The MSE is 0.0487, indicating the very good forecasting ability of the wavelet network. On the other hand, MAPE and SMAPE are relatively high, presenting the high impact of the large noise term.

TABLE 6.6 Case 1: Out-of-Sample Error Criteria[a]

Md.AE	MAE	MaxAE	SSE	RMSE	NMSE	MSE	MAPE	SMAPE
0.1243	0.1669	0.7171	14.6254	0.2208	0.3422	0.0487	145.67%	62.59%

[a]Md.AE, median absolute error; MAE, mean absolute error; MaxAE, maximum absolute error; SSE, sum of squared errors; RMSE, root mean squared error; NMSE, normalized mean squared error; MSE, mean squared error; MAPE, mean absolute percentage error; SPAME, symmetric mean absolute percentage error.

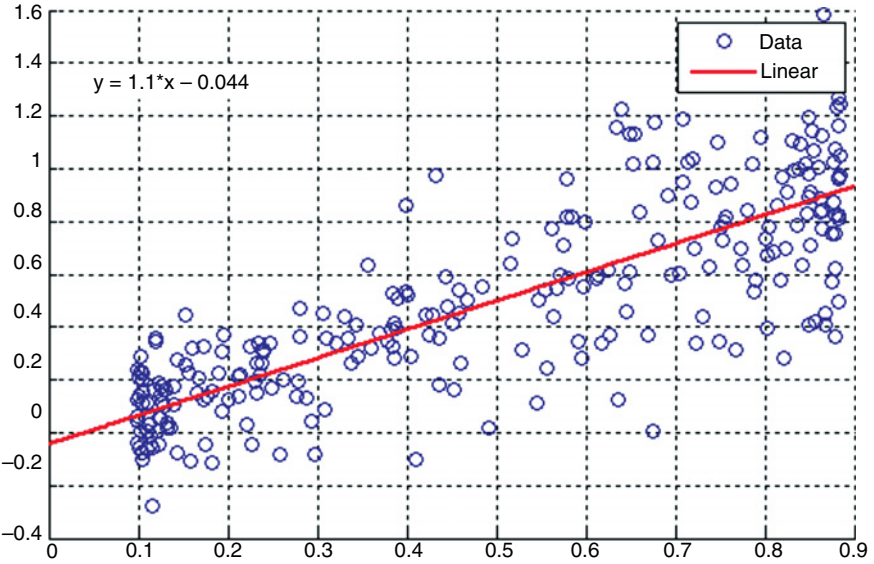

Figure 6.4 *Case 1: out-of-sample scatter plot.*

The scatter plot of the out-of-sample forecasts and the target values is presented in Figure 6.4, and the estimated parameters of the linear regression between the values forecasted and the target values are presented in Table 6.7. A close inspection of the table reveals that the parameter b_0 is not statistically different from zero, while the parameter b_1 is not statistically different from 1 at a confidence level of 5%. Moreover, the linear regression is statistically significant according to the F-statistic. Finally, as presented in Table 6.8, POCID, IPOCID, and POS are 65.9%, 45.5%, and 92.3%, respectively.

TABLE 6.7 Case 1: Out-of-Sample Regression Statistics[a]

	Parameter	p-Values	S.E.	T-Stat.
b_0	−0.044029	0.082175	0.025245	−1.744089
b_1	1.08430	0.000000	0.044898	24.150415
$b_1 = 1$ test	1.08430	0.061413	0.044898	1.877597
R^2	0.661841			
F	583.243	0.000000		
DW	1.870093	0.235214		

[a]S.E., squared error; DW, Durbin–Watson.

TABLE 6.8 Case 1: Change-in-Direction Metrics[a]

POCID	IPOCID	POS
65.9%	45.5%	92.3%

[a]POCID, prediction of change in direction; IPOCID, independent prediction of change in direction; POS, prediction of sign.

TABLE 6.9 Case 2: Residuals Testing[a]

	Parameter	p-Values
n/p Ratio	38.5000	
Mean	0.0002	
Median	0.0004	
S. dev.	0.0661	
DW	1.9415	0.3540
LB Q-stat.	10.9627	0.9472
JB stat.	908,651.8641	0.0000
KS stat.	14.4117	0.0000
R^2	0.9974	
$\bar{R}^2$	0.9973	

[a]S.dev., standard deviation; DW, Durbin–Watson; LB, Ljung–Box; KS, Kolmogorov–Smirnov.

Case 2: Sum of Sinusoids and Cauchy Noise

As discussed in Chapters 3 and 4, a wavelet network with 8 hidden units was used. This simulated case incorporated large outliers to the data. In this section we examine how well the wavelet network was trained on the data. In other words, we assess the ability of the network to learn the data and then forecast the future values of the underlying function.

Testing the Residuals First, the residuals of the trained networks are examined. The various descriptive statistics as well as various tests on the residuals are presented in Table 6.9. The mean of the residuals is close to zero, 0.002, while the standard deviation is only 0.0661. The Ljung–Box tests reject the hypothesis of correlation on the residuals. The Komlogorov–Smirnov and the Jarque–Berra statistics indicate the absence of normality in the residuals, which is logical given the fact the residuals were generated by a Cauchy process. Finally, both the R^2 and the $\bar{R}^2$ show a good fit of the network to the data, with values over 99.7%.

Error Criteria Various error criteria are presented in Table 6.10. These criteria will help us assess the fit of the wavelet network to the data: in other words, how well the wavelet network learned the data. A closer inspection of the table reveals that all criteria are very small, indicating a very good fit of the wavelet network to the data. The MSE and NMSE 0.0044 and 0.0026, respectively. On the other hand, MAPE and SMAPE are only 3.80% and 2.03%, respectively.

TABLE 6.10 Case 2: Error Criteria[a]

Md.AE	MAE	MaxAE	SSE	RMSE	NMSE	MSE	MAPE	SMAPE
0.0084	0.0201	1.1068	4.3683	0.0661	0.0026	0.0044	3.80%	2.03%

[a]Md.AE, median absolute error; MAE, mean absolute error; MaxAE, maximum absolute error; SSE, sum of squared errors; RMSE, root mean squared error; NMSE, normalized mean squared error; MSE, mean squared error; MAPE, mean absolute percentage error; SPAME, symmetric mean absolute percentage error.

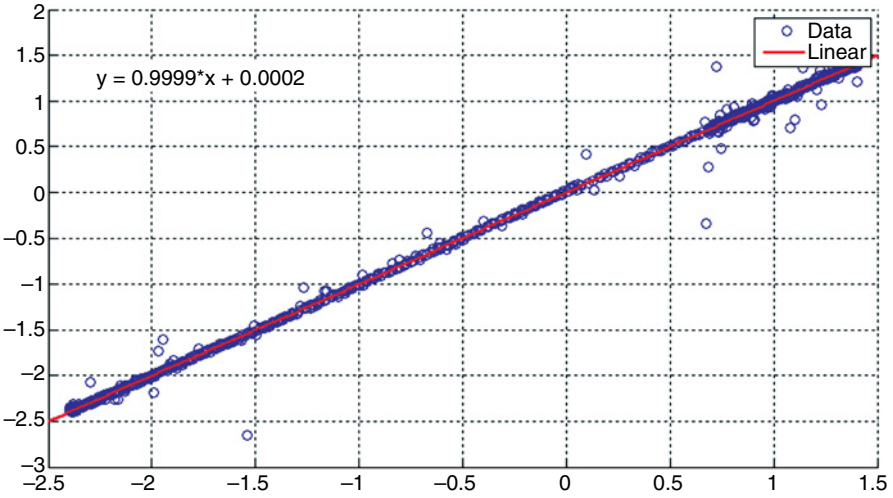

Figure 6.5 *Scatter plot for the second simulated case.*

Scatter Plot Figure 6.5 is a scatter plot of the real values versus the values forecasted for the wavelet network. As can be seen, the forecasts and the targets are in line.

Regression Between the Forecasts and the Target Values The estimated parameters of the regression between the forecasted values and the target values are presented in Table 6.11. A close inspection reveals that the parameter b_0 is not statistically different from zero, while the parameter b_1 is not statistically different from 1. Moreover, the linear regression is statistically significant according to the F-statistic.

Changes in Direction The POCID, IPOCID, and POS criteria are presented in Table 6.12. In this application the aim for the wavelet network is to learn the underlying process in the presence of large outliers. The POCID, IPOCID, and POS are 81.40%, 75.10%, and 99.80%, indicating that the wavelet network can successfully identify and predict the changes in the direction of the underlying function.

TABLE 6.11 Case 2: Regression Statistics for the Second Simulated Case[a]

	Parameter	p-Values	S.E.	T-Stat.
b_0	0.0002	0.9147	0.0021	0.1072
b_1	0.9999	0.0000	0.0016	619.8138
$b_1 = 1$ test	0.9999	0.9577	0.0016	−0.0531
R^2	0.9974			
F	384,170.0000	0.0000		
DW	1.9415	0.338		

[a]S.E., squared error; DW, Durbin–Watson.

TABLE 6.12 Case 2: Change-in-Direction Metrics[a]

POCID	IPOCID	POS
81.40%	75.10%	99.80%

[a]POCID, prediction of change in direction; IPOCID, independent prediction of change in direction; POS, prediction of sign.

TABLE 6.13 Case 2: Out-of-Sample Residuals Testing[a]

	Parameter	*p*-Values
n/p Ratio	11.5385	
Mean	0.0043	
Median	−0.0008	
S. dev.	0.0655	
DW	2.0068	0.9523
LB *Q*-stat.	15.3770	0.7544
JB stat.	143,397.8894	0.0000
KS stat.	7.9658	0.0000
R^2	0.9971	
$\bar{R}^2$	0.9969	

[a]S. dev., standard deviation; DW, Durbin–Watson; LB, Ljung–Box; KS, Kolmogorov–Smirnov.

Out-of-Sample Forecasts In this section the out-of-sample performance of the wavelet network is evaluated. The various descriptive statistics as well as various tests on the out-of-sample residuals are presented in Table 6.13. The mean of the residuals is close to zero with a standard deviation of 0.06. The Ljung–Box test accepts the hypothesis of uncorrelated residuals. Similarly, the Komlogorov–Smirnov and Jarque–Berra tests indicate the absence of normality in the residuals. Finally, both R^2 and $\bar{R}^2$ are over 99.6%, showing the good predictive ability of the network.

Next, the accuracy of the prediction is evaluated. Various error criteria are presented in Table 6.14. The values of the metrics are similar to those presented in Table 6.10. MSE, NMSE, and RMSE are 0.0043, 0.0029, and 0.0656, indicating the very good forecasting ability of the wavelet network. On the other hand, MAPE and SMAPE are only 3.88% and 0.19%.

Figure 6.6 is a scatter plot of the out-of-sample forecasts and the target values. The estimated parameters of the linear regression between the values forecasted and the target values are presented in Table 6.15. A close inspection of the table

TABLE 6.14 Case 2: Out-of-Sample Error Criteria[a]

Md.AE	MAE	MaxAE	SSE	RMSE	NMSE	MSE	MAPE	SMAPE
0.0074	0.0198	0.8572	1.2903	0.0656	0.0029	0.0043	3.88%	0.19%

[a]Md.AE, median absolute error; MAE, mean absolute error; MaxAE, maximum absolute error; SSE, sum of squared errors; RMSE, root mean squared error; NMSE, normalized mean squared error; MSE, mean squared error; MAPE, mean absolute percentage error; SPAME, symmetric mean absolute percentage error.

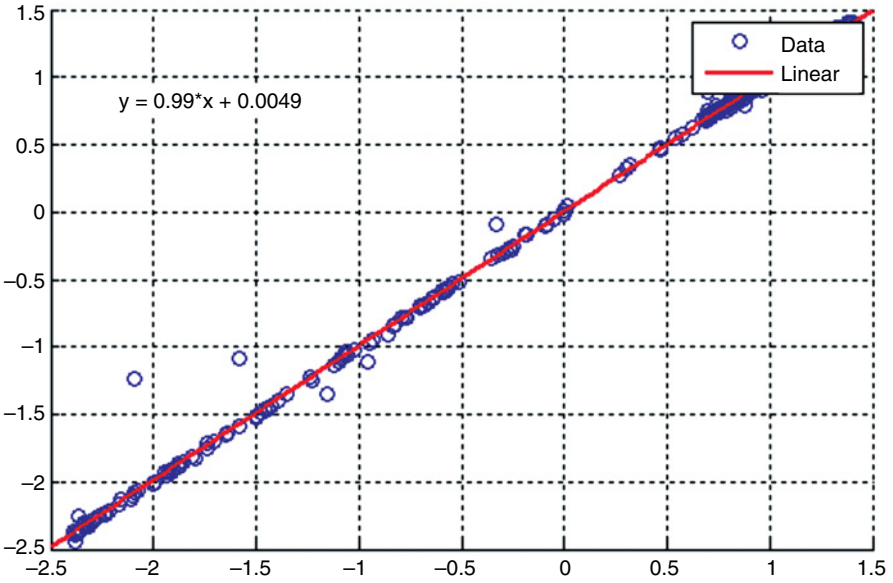

Figure 6.6 *Case 2: out-of-sample scatter plot.*

TABLE 6.15 Case 2: Out-of-Sample Regression Statistics[a]

	Parameter	*p*-Values	S.E.	*T*-Stat.
b_0	0.0049	0.1928	0.0038	1.3052
b_1	0.9944	0.0000	0.0031	324.9527
$b_1 = 1$ test	0.9944	0.0664	0.0031	−1.8425
R^2	0.9972			
F	105,594.2273	0.0000		
DW	2.0384	0.7830		

[a]S.E., squared error; DW, Durbin–Watson.

reveals that the parameter b_0 is not statistically different from zero and the parameter b_1 is not statistically different from 1 at a confidence level of 5%. Moreover, the linear regression is statistically significant according to the F-statistic. Finally, as presented in Table 6.16, POCID, IPOCID, and POS are 82.29%, 79.93%, and 99.33%, respectively.

TABLE 6.16 Case 2: Change-in-Direction Metrics[a]

POCID	IPOCID	POS
86.29%	79.93%	99.33%

[a]POCID, prediction of change in direction; IPOCID, independent prediction of change in direction; POS, prediction of sign.

CLASSIFICATION

In classification applications, the primary objective is to determine the group to which an "object" (person, company, product, etc.) belongs (Green and Carroll, 1978). In finance, some of the usual classification applications are: prediction of the success or failure of a new product, determination of credit risk of client, and bankruptcy probability.

In statistical terminology such applications are known as *applications of discriminant analysis*. Discriminant analysis is the appropriate statistical approach when the dependent variable is categorical—nominal or nonmetric—while the independent variables are numerical.

If the dependent variable is composed of two groups or classifications, it is called *discriminant analysis of two groups*. When the dependent variable is composed of three or more groups or classifications, it is called *multiple discriminant analysis*.

Discriminant analysis involves extraction of the linear combination of dependent variables that will better distinguish among the a priori fixed groups. The linear combinations for a discriminant analysis are derived from an equation that takes the following form:

$$z = w_1 x_1 + w_2 x_2 + \cdots + w_m x_m \tag{6.47}$$

where $i = 1, \ldots, m$, x_i are the independent variables, w_i are the discriminant weights, and z is the discriminant score. In the case of wavelet networks, the linear relationship (6.47) is replaced by the following nonparametric relationship:

$$z = g_\lambda \left(\mathbf{x}; \hat{\mathbf{w}}_n \right) \tag{6.48}$$

From equation (6.47) it can be derived that in discrete analysis each independent variable is multiplied by its respective weight and then the products are added to estimate the discrete score of vector x_i. The average value of the discrete scores of all individuals within a given group is called a *centroid*. The number of centroids is equal to the number of groups. For simplicity, we restrict our analysis to applications with only two centroids. The greater the distance between the centroids, the smaller the overlap of the distributions of distinct scores of the two groups and therefore the better the distinct function.

A distinct function that provides good separation between classes A and B is presented in Figure 6.7. On the other hand, a distinct function that performs a relatively poor separation between classes A and B is presented in Figure 6.8.

Assumptions and Objectives of Discriminant Analysis

The basic assumptions under the (linear) discriminant analysis is a multivariate normality of the independent variables and unknown (but equal) dispersion and covariance matrices for the groups as defined by the dependent variable (Green and Carroll, 1978; Harris, 2001).

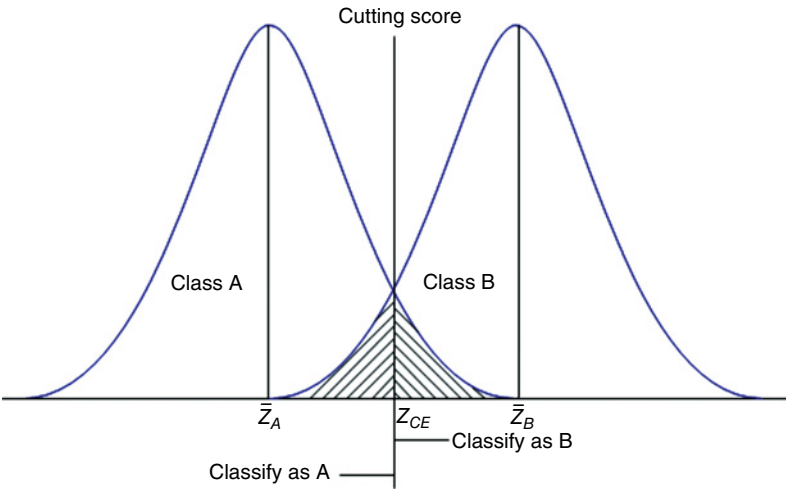

Figure 6.7 *Good separation between groups A and B and the optimal cutoff value for samples of the same size.*

Additionally, it must be ensured that the independent variables are truly linearly independent and that multicollinearity between the input variables does not exist. This assumption becomes especially important when stepwise variable selection methods, such as the ones presented in Chapter 5, are used (Hair et al., 2010).

Finally, regarding the size of the sample (either of the training or of the validation sample), a general rule that many studies suggest is that the sample must contain at least 20 observation for each independent variable (Hair et al., 2010).

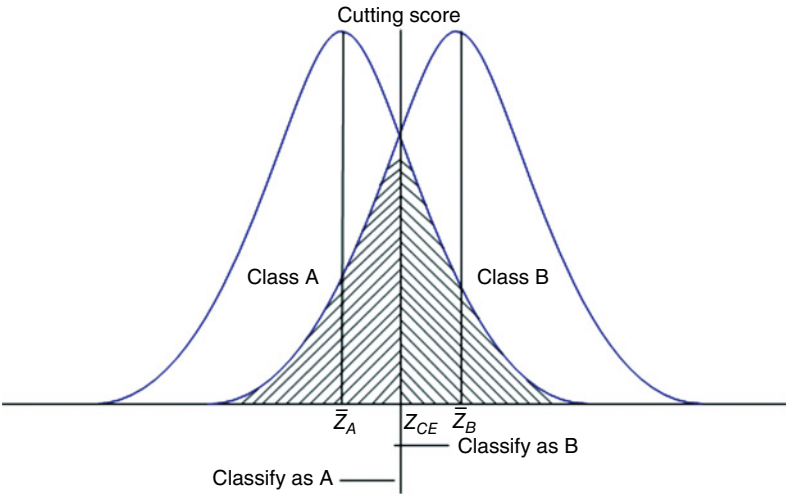

Figure 6.8 *Insufficient separation between groups A and B and the cutoff value for samples of the same size.*

The objectives of discriminant analysis are summarized as follows:

- To create models for the classification of individuals or objects into groups on the basis of their scores on several variables.
- To determine whether statistically significant differences exist between the average score profiles of the two (or more) a priori defined groups.
- To determine the independent variables that contribute most in the average discriminant score.

It is clear from the foregoing objectives that discriminant analysis is useful in the correct classification of statistical units into groups or classes or in an understanding of the differences among the various groups. Hence, discriminant analysis can be used either as a predictive technique or as a type of profile analysis (Hair et al., 2010). In this book we are interested in using discriminant analysis as a predictive technique.

Validation of the Discriminant Function

The next step in discriminant analysis is the validation. After estimation of the discriminant function, the statistical significance of the function must be estimated. This can be done either by the calculation of chi-square or the Mahalanobis D^2 statistics. Although these statistics assess the significance of the discriminant function, they do not provide any information about the predictive power of the function.

On the other hand, the predictive power of the discriminant function can be determined by the classification matrix. The classification matrix can be related to the concept of the R^2 in linear regression. As in the case of linear regression, there are cases where the regression is statistically significant; however, the R^2 is very low. In other words, the linear regression explains a very small percentage of the variance. Similarly, the discriminant function can be statistically significant but with low predictive ability. In discriminant analysis the percentage of the cases classified correctly is called the *hit ratio* and is analogous to R^2.

Cutting Score Determination Before proceeding with construction of the classification matrix, we need to determine the cutting score. The *cutting score criterion* is the score against which each case's discriminant score is compared to determine into which group the observation should be classified. To construct the classification matrix, the optimal cutting score or critical Z values must be determined. The optimal cutting score will differ depending on whether the sizes of the groups are equal or unequal.

If the two groups are of equal size, that is, have the same number of observations, the optimal cutting score will be located in the middle of the distance between the two centroids. In this case the cutting score is given by

$$z = \frac{\bar{z}_A + \bar{z}_B}{2} \tag{6.49}$$

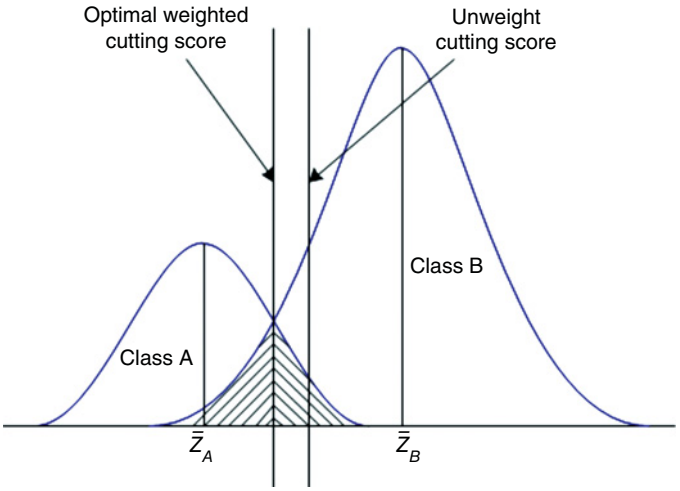

Figure 6.9 *Cutting score for two classes of unequal sample size.*

where z is the critical cutting score value for equal group sizes, $\bar{z}_A$ is the centroid for group A, and $\bar{z}_B$ is the centroid for group B.

If the two groups are not of equal size, the cutting score is estimated by a weighted average as follows:

$$z = \frac{n_A \bar{z}_A + n_B \bar{z}_B}{n_A + n_B} \tag{6.50}$$

where z is the critical cutting score value for unequal group sizes, $\bar{z}_A$ is the centroid for group A, $\bar{z}_B$ is the centroid for group B, n_A is the number of observations in group A, and n_B is the number of observations in group B. An example of a cutting score for two groups of unequal size is presented in Figure 6.9. Note that formula (6.49) is a special case of (6.50) with $n_A = n_B$. Furthermore, in formulas (6.49) and (6.50), a normal distribution and a known covariance are assumed.

In determining the optimal cutoff score, the cost of classifying a case incorrectly should be taken into account. If the costs of an incorrect prediction in various cases are about the same, the optimal cutoff score will be the one that will classify incorrectly the smaller number of cases in all groups. If the costs of misclassifications are not the same, the optimal cutoff score will be the one that minimizes the total cost of the misclassification. For an in-depth analysis in defining the optimum cutoff score, we refer to Dillon and Goldstein (1984) and Huberty et al. (1987).

For construction of the classification matrix, for each case in which it is included in the testing sample, its discriminant score must be compared against the cutting score. Then each case is classified as follows:

- Classify case n intro group A if $z_n < z_{cs}$.
- Classify case n intro group B if $z_n > z_{cs}$.

TABLE 6.17 Classification Matrix

	Forecast		
Target	0	1	Total
0	α	β	$\alpha + \beta$
1	γ	δ	$\gamma + \delta$
Total	$\alpha + \gamma$	$\beta + \delta$	$\alpha + \beta + \gamma + \delta$

Here z_n is the discriminant score of observation n and z_{cs} is the critical cutting score value.

The results of the classification are represented in matrix form. Table 6.17 shows an example of a classification matrix.

The significance of the classification accuracy can be determined by the following t-test in the case of two equal sample groups:

$$t = \frac{p - 0.5}{\sqrt{0.5\,(1 - 0.5)/n}} \tag{6.51}$$

where p is the proportion classified correctly and n is the sample size. Formula (6.51) can easily be modified to be used with more groups and unequal sample sizes.

Evaluating the Classification Ability of a Wavelet Network

In the preceding section, the classification accuracy of a discriminant function was measured by the hit ratio, which is obtained by the classification matrix. Additional measures that can describe the predictive and classification accuracy are presented in this section.

Relative Entropy The first measure presented, the relative entropy, can be estimated as follows:

$$\text{RE} = -\frac{1}{n}\left[\sum_{s=1}^{n_A} \ln\left(1 - y_{\text{A}}^{(s)}\right) + \sum_{t=1}^{n_B} \ln\left(y_{\text{B}}^{(t)}\right)\right] \tag{6.52}$$

where n_{A} and n_{B} are the observations in class A (0) and B (1), respectively, and $y_{\text{A}}^{(s)}$ and $y_{\text{B}}^{(t)}$ are the outputs of the network for the patterns s in class 0 and t in class 1. If the relative entropy is close to zero, it indicates a good fit to the data.

Kolmogorov–Smirnov Statistic Another indicator of the classification power of a wavelet network that can be used is the Kolmogorov–Smirnov (KS) statistic. To estimate the KS statistic, cumulative histograms of the projected outputs of the network are computed and their overlaps are estimated.

The KS statistic is the maximum distance between the respective percentile values of the two histograms. If the output values of the model were random, the cumulative

histograms would both be linear and the KS statistic would be close to zero. If the model were perfect, the KS statistic would be equal to 1.

Maximum Chance Criterion Before we proceed with our analysis, we must define what is considered an acceptable level of prediction for a distinct function. This depends on the percentage number of observations that can be classified correctly in a random manner. If every class consists of the same number of observations, the percentage is 1 divided by the number of classes.

However, when the number of observations is different for each class, the maximum chance criterion is used. This criterion is determined by the largest class. For example, if the data set consists of two classes, where the first has 65 observations and the second has 35, the maximum chance criterion is 65%. In other words, if our model cannot classify correctly at least 65% of the observations, it does not have any classification power.

Proportional Chance Criterion The preceding criterion should be used when the sole objective aim of the analysis is to maximize the proportion of cases classified correctly. However, we are usually interested in the correct classification in both classes. In other words, the criterion above does not take into account the classification in the smaller group. Returning to the preceding example, where class A (0) had 35 observations and class B (1) had 65 observation. If we classify all cases (all of the 100 cases) as 1, our strategy will produce 65 correct classifications.

Alternatively, the proportional chance criterion can be used:

$$\text{PRO} = p^2 + (1 - p)^2 \tag{6.53}$$

where p is the percentage of the cases that belong to class A and $1 - p$ is the percentage of the cases that belong to class B.

Classification Accuracy Relative to Chance A crucial question that must be asked is whether our model has a percentage of correct classification significantly larger than would be expected by chance. If the wavelet network has classification accuracy greater than the one that can be expected by chance, we can proceed on the interpretation of the discriminant functions. Otherwise, an analysis would not provide meaningful results (Hair et al., 2010).

The following quick criterion is presented by Hair et al. (2010): The classification accuracy should be at least 25% greater than that achieved by chance. For example, if chance accuracy is 50%, the classification accuracy should be at least 62.5%.

A more robust test is Press's Q-statistic. This measure compares the number of correct classifications with the total sample size and the number of groups. The value calculated is then compared with a critical value obtained for chi-square with 1 degree of freedom. If the estimated value is greater than the critical value, a significantly better classification ability can be obtained using the model than by chance.

Press's Q-statistic is given by

$$Q = \frac{[n - (c \times k)]^2}{n\,(k - 1)} \tag{6.54}$$

where n is the sample size, c the number of correct classifications, and k the number of classification groups. This test is sensitive to sample size, where large samples are more likely to show significance than small sample sizes of the same classification rate.

Sensitivity The sensitivity test measures the ability of a model to identify positive results. The sensitivity, also called the *true positive rate*, is given by the relationship

$$\text{sensitivity} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} = \frac{\text{true positives}}{\text{total positives}} \tag{6.55}$$

Specificity Similarly, the specificity test measures the ability of a model to identify negative results. The specificity, also called the *true negative rate*, is given by the following relationship:

$$\text{specificity} = \frac{\text{true negatives}}{\text{true negatives} + \text{false positives}} = \frac{\text{true negatives}}{\text{total negatives}} \tag{6.56}$$

Rate of Correctness (Hit Ratio) As already presented, the rate of correctness or hit ratio measures the percentage of the correct classification of the wavelet network:

$$\text{RC} = \frac{\text{true positives} + \text{true negatives}}{\text{true positives} + \text{true negatives} + \text{false positives} + \text{false negatives}} \tag{6.57}$$

Rate of Missing Chances The rate of missing chance measures the percentage of true cases that were classified as false:

$$\text{RMC} = \frac{\text{false negatives}}{\text{false negatives} + \text{true positives}} \tag{6.58}$$

Rate of Failure The rate of failure measures the percentage of false cases identified as true.

$$\text{RF} = \frac{\text{false positives}}{\text{false positives} + \text{true negatives}} \tag{6.59}$$

Fitness Functions The criteria above can be used to construct fitness functions. Fitness functions are useful for comparing different models. An example of a fitness function is

$$\text{fitness} = 0.6\,\text{RC} - 0.1\,\text{RMC} - 0.3\,\text{RF} \tag{6.60}$$

In equation (6.60), different weights can be used, depending on the application.

Case 3: Classification Example on Bankruptcy

In this section a different problem is considered. A wavelet network is constructed to classify if a firm will or will not go bankrupt based on various attributes. The data set contains samples from 240 Greek firms. Each instance has one of two possible classes: bankrupt or nonbankrupt. The two groups have the same number of firms; hence, there are 120 firms in each group. The objective is to construct a wavelet network that accurately classifies each firm. The classification is based on four attributes proposed by Altman and Saunders (1997):

$$X_1 = \frac{\text{working capital}}{\text{total assets}}$$

$$X_2 = \frac{\text{retained earnings}}{\text{total assets}}$$

$$X_3 = \frac{\text{earnings before interest and tax}}{\text{total assets}}$$

$$X_4 = \frac{\text{book value of equity}}{\text{total liabilities}}$$

Our results will be compared against a linear classification model.

The data were split randomly into training and validation samples. The training sample consists of 168 (70%) cases. The validation sample consists of 72 (30%) cases and is used to evaluate the predictive and classification power of the trained wavelet network.

In the training sample, 86 firms went bankrupt and were given the value 0, and 82 firms that were not bankrupt and were given the value 1. The cutting score is 0.4882. A classification matrix of the training sample using a wavelet network is presented in Table 6.18. A closer inspection of the table reveals very good classification rates. More specifically, the wavelet network classified correctly 56 nonbankrupt firms and 70 bankrupt cases. Hence, the wavelet network classified correctly 126 of 168 cases (75%). The specificity of the models is 68.29% and the sensitivity is 81.40%. On the other hand, the rate of failure and the rate of missing chances are 18.60% and 31.71%, respectively. Finally, the fitness function given by (6.60) is 0.3625.

Similarly, in Table 6.19 the classification matrix is presented when a linear model is used. A closer inspection reveals that the classification ability of the linear model in-sample is worse. The hit ratio is only 70.83% and the sensitivity and specificity

TABLE 6.18 In-Sample Classification Matrix: Wavelet Network

	Forecast				
Target	Nonbankrupt	Bankrupt	Total	Sensitivity	Specificity
Nonbankrupt	56	26	82	68.29%	81.40%
Bankrupt	16	70	86	Rate of Missing Chances	Rate of Failure
Total	72	96	168	31.71%	18.60%

TABLE 6.19 In-Sample Classification Matrix: Linear Case

	Forecast				
Target	Bankrupt	Nonbankrupt	Total	Sensitivity	Specificity
Bankrupt	55	27	82	67.07%	74.42%
Nonbankrupt	22	64	86	Rate of Missing Chances	Rate of Failure
Total	77	91	168	32.93	25.58%

TABLE 6.20 Evaluation of the Classification Ability of the Wavelet Network

Maximum Chance	1.25% Max. Chance	Pro	Press's Q^a	Hit Ratio
51.19%	63.99%	50.03%	42	75.00%

[a]Press's Q critical values at confidence levels 0.1, 0.05, and 0.01: 2.71, 3.84, 6.63.

are 67.07% and 74.42%, respectively. Finally, the fitness function is reduced and is only 0.3153.

The evaluation of the classification ability of the wavelet network is presented in Table 6.20. The maximum chance criterion is 51.19% and the heuristic method presented by Hair et al. (2010) is 63.99%. Finally, the proportional chance criterion is 51.77%. The hit ratio is significantly larger than the various chance criteria and is 75%. Hence, the model is predicting significantly better than chance. This is also confirmed by the large value of Press's Q-statistic, which is greater than the critical values at confidence levels 0.1, 0.05, and 0.01. Similarly, the evaluation of the classification ability of the linear model is presented in Table 6.21. The hit ratio is 70.83% and Press's Q-statistic is higher than the critical values. Hence, both the wavelet network and the linear model predict significantly better than chance.

Out-of-Sample Next, the forecasting and classification ability of the trained wavelet network is evaluated in the validation sample. The data of the validation sample were not used during the training phase. Hence, these are new data that were never presented to the wavelet network.

In the validation sample there are 40 bankrupt cases given the value 0 and 32 nonbankrupt cases given the value 1. The classification matrix of the training sample using the wavelet network is presented in Table 6.22. A closer inspection of the table reveals very good prediction ability and classification rates. More specifically, the wavelet network classified correctly 24 nonbankrupt cases and 31 bankrupt cases. Hence, the wavelet network classified correctly 55 of 72 cases (76.39%). The specificity of the models is 77.50% and the sensitivity is 75%. On the other hand, the rate

TABLE 6.21 Evaluation of the Classification Ability of the Linear Model

Maximum Chance	1.25% Max. Chance	Pro	Press's Q^a	Hit Ratio
51.19%	63.99%	50.03%	29.16	70.83%

[a]Press's Q critical values at confidence levels 0.1, 0.05, and 0.01: 2.71, 3.84, 6.63.

TABLE 6.22 Out-of-Sample Classification Matrix: Wavelet Network

	Forecast				
Target	Nonbankrupt	Bankrupt	Total	Sensitivity	Specificity
Nonbankrupt	24	8	32	75.00%	77.50%
Bankrupt	9	31	40	Rate of Missing Chances	Rate of Failure
Total	33	39	72	25.00%	22.50%

TABLE 6.23 Out-of-Sample Classification Matrix: Linear Case

	Forecast				
Target	Nonbankrupt	Bankrupt	Total	Sensitivity	Specificity
Nonbankrupt	23	9	32	71.88%	65.00%
Bankrupt	14	26	40	Rate of Missing Chances	Rate of Failure
Total	37	35	72	28.12%	35.00%

of failure and the rate of missing chances are low, 22.50% and 25%, respectively. Finally, the fitness function is 0.3658.

The out-of-sample classification matrix of the linear model is presented in Table 6.23. It is clear that the classification accuracy is reduced when a linear model is used. More precisely, the sensitivity and specificity were reduced to 71.88% and 65% while the linear model classified correctly only 49 of the 72 cases. Similarly, the rate of missing chances and the rate of failure were increased to 28.12% and 35%. Finally, the fitness function is 0.2752.

The evaluation of the classification ability of the wavelet network and the linear model are presented in Tables 6.24 and 6.25, respectively. The maximum chance criterion is 55.56% and the heuristic method presented by Hair et al. (2010) is 69.44%. Finally, the proportional chance criterion is 50.62%. The hit ratio of the wavelet network is significantly larger than the maximum chance and the proportional chance criteria and is 76.39%. Similarly, the hit ratio of the linear model is 68.06%. Hence, both the linear and the wavelet network models are predicting significantly better than chance. This is confirmed by the large value of Press's Q-statistic, which

TABLE 6.24 Out-of-Sample Evaluation of the Classification Ability of the Wavelet Network

Maximum Chance	1.25% Max. Chance	Pro	Press's Q^a	Hit Ratio
55.56%	69.44%	50.62%	20.05	76.39%

[a]Press's Q critical values at confidence levels 0.1, 0.05, and 0.01: 2.71, 3.84, 6.63.

TABLE 6.25 Out-of-Sample Evaluation of the Classification Ability of the Linear Model

Maximum Chance	1.25% Max. Chance	Pro	Press's Q^a	Hit Ratio
55.56%	69.44%	50.62%	9.39	68.06%

[a]Press's Q critical values at condidence levels 0.1, 0.05, and 0.01: 2.71, 3.84, 6.63.

is greater than the critical values at confidence levels 0.1, 0.05, and 0.01. However, it is clear that the wavelet network outperforms the linear model. Not only is the hit ratio significantly higher when a wavelet network is used, but the rate of failure, which leads to false decisions and loss of money, is also significantly smaller.

CONCLUSIONS

In this chapter various metrics for measuring model adequacy and the predicting ability of a wavelet network were presented. First, various tests were presented that test the properties of the residuals of the fitted wavelet network. Next, the forecasting ability of a wavelet network was tested using scatter plots and a linear regression between the values forecasted obtained from the wavelet network and the target values. Depending on the application, we may be interested in three categories of predictions. In the first case, we are interested in value forecasting. In this case, the predictive power of the wavelet network is measured as an expression of the difference between the target values and the output of the network. Second, in classification applications we are interested in the correct classification of various cases. In this case the predictive power of the network is measured as the ability of the wavelet network to classify the individual cases correctly. Finally, there are applications where we are interested only in the sign or in the change in the direction of the values. In this case, metrics such as the POS and IPOCID are used.

In this section the wavelet network models were tested for adequacy in three cases. The first two were time-series approximation and forecasting problems; the last was a classification problem. More precisely, in the last case study, a wavelet network was constructed and used to classify the credit risk of firms. Our results indicate that the nonlinear nonparametric wavelet network outperforms the linear model significantly.

REFERENCES

Altman, E. I., and Saunders, A. (1997). "Credit risk measurement: developments over the last 20 years." *Journal of Banking and Finance*, 21(11–12), 1721–1742.

Box, G. E. P., and Jenkins, G. M. (1970). *Time Series Analysis, Forecasting and Control*, Holden-Day, San Francisco.

Dillon, W. R., and Goldstein, M. (1984). *Multivariate Analysis*. Wiley, New York.

Durbin, J., and Watson, G. S. (1951). "Testing for serial correlation in least squares regression: II." *Biometrika*, 38(1/2), 159–177.

Green, P. E., and Carroll, J. D. (1978). *Analyzing Multivariate Data*. Dryden Press, Hinsdale, IL.

Hair, J., Black, B., Babin, B., and Anderson, R. (2010). *Multivariate Data Analysis*, 7th ed. Prentice Hall, Upper Saddle River, NJ.

Harris, R. J. (2001). *A Primer of Multivariate Statistics*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Huberty, C. J., Wisenbaker, J. M., and Smith, J. C. (1987). "Assessing predictive accuracy in discriminant analysis." *Multivariate Behavioral Research*, 22(3), 307–329.

Judge, G., Hill, R. C., Griffiths, W. E., Lutkepohl, H., and Lee, T.-C. (1982). *Introduction to the Theory and Practice of Econometrics*. Wiley, New York, p. 880.

Makridakis, S., Wheelwright, S. C., and Hyndman, R. J. (2008). *Forecasting Methods and Applications*. Wiley, Hoboken, NJ.

Ramsey, J. B. (1969). "Tests for specification errors in classical linear least-squares regression analysis." *Journal of the Royal Statistical Society*, Seri B, 350–371.

White, H. (1989). "Learning in artificial neural networks: a statistical perspective." *Neural Computation*, 1, 425–464.

Yao, J., Tan, C. L., and Poh, H.-L. (1999). "Neural networks for technical analysis: a study on KLCI." *International Journal of Theoretical and Applied Finance*, 2(02), 221–241.

Zapranis, A., and Refenes, A. P. (1999). *Principles of Neural Model Indentification, Selection and Adequacy: With Applications to Financial Econometrics*, Springer-Verlag, New York.