5

Variable Selection: Determining the Explanatory Variables

As mentioned in earlier chapters, the model identification procedure is divided into two parts: model selection and variable selection. In this chapter we focus on the second part of the model identification procedure, variable selection.

The use of absolutely necessary explanatory variables is known as the *principle of parsimony*, or *Occam's razor*. In real problems it is important for various reasons to determine correctly the independent variables. In most financial problems there is little theoretical guidance and information about the relationship of any explanatory variable with the dependent variable. As a result, unnecessary independent variables are included in the model, reducing its predictive power. The use of irrelevant variables is among the most common sources of specification error. Also, correctly specified models are easier to understand and to interpret. The underlying relationship is explained by only a few key variables, while all minor and random influences are attributed to the error term. Finally, including a large number of independent variables relative to sample size (an *overdetermined* model) runs the risk of a spurious fit.

To select the statistical significant and relevant variables from a group of possible explanatory variables, an approach involving the significance of statistical tests of hypotheses is followed. To do so, the relative contributions of the explanatory variables in explaining the dependent variable in the context of a specific model are estimated. Then the *significance* of a variable is assessed. This can be done by testing the null hypothesis that it is "irrelevant," either by comparing a test statistic with a known theoretical distribution with its critical value or by constructing confidence

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

© 2014 John Wiley & Sons, Inc. Published 2014 by John Wiley & Sons, Inc.

intervals for the relevance criterion. Our decision as to whether or not to reject the null hypothesis is based on a given significance level. The p-value, the smallest significance level for which the null hypothesis will not be refuted, imposes a ranking on the variables according to their relevance to the particular model (Zapranis and Refenes, 1999).

Evaluating the statistical significance of explanatory variables involves the following three stages:

- Defining a variable's "relevance" to the model
- Estimating the sampling variation of the relevance measure
- Testing the hypothesis that the variable is "irrelevant"

Before proceeding to the variable selection, a measure of relevance must be defined. For nonlinear models the derivative $\partial y / \partial x_j$ is not constant. As a result, new composite measures of the sensitivity of y on x_j were developed. Practitioners usually employ measures such as the average derivative and the average derivative magnitude. In the next sections, various sensitivity measures are presented. Alternatively, model-fitness sensitivity (MFS) criteria can be used. The model-fitness sensitivity criteria quantify the effect of the explanatory variable on the empirical loss and on the coefficient of determination, R^2.

Estimating sampling variability is an important part of this chapter. Without knowledge of the sampling distributions of the relevance measures, there is no way of knowing to which level the estimates available are affected by random sampling variation. In this chapter the bootstrapping and cross-validation methods are used to estimate the sampling distributions and to compute the p-values for variable significance hypothesis testing.

In this chapter we also outline a framework for hypothesis testing using nonparametric confidence intervals for the sensitivity and the MFS criteria in the context of variable selection. In this section, various methods of testing the significance of each explanatory variable are presented and tested. The purpose of this section is to find an algorithm that constantly gives stable and correct results when it is used with wavelet networks.

Once a variable is removed as being insignificant, the correctness of that action has to be evaluated. It is not uncommon, for example, to overestimate standard error in the presence of multicollinearity. A usual approach is to compare the reduced model with the full model on the basis of some performance criterion. For this purpose the prediction risk is used. Moreover, the adjusted coefficient of determination for degrees of freedom, $\bar{R}^2$, is computed, and we outline an iterative variable selection process with backward variable elimination. Finally, we use two case studies to demonstrate the framework we propose for variable selection.

EXISTING ALGORITHMS

In linear models the coefficient of an explanatory variable reflects the reactions of the dependent variable to small changes in the value of the explanatory variable.

However, the value of the coefficient does not provide any information about the significance of the corresponding explanatory variable. Hence, in linear models, to determine if a coefficient, and as a result an input variable, are significant, the t-statistics or p-values of each coefficient are examined. Applying such a method in wavelet networks is not a straightforward process since the coefficients (weights) are estimated iteratively and each variable contributes to the output of the wavelet network linearly through the direct connections and nonlinearly through the hidden units. Moreover, the finite-sample distribution of the network parameters is not known either; it must be estimated empirically, or asymptotic arguments have to be used.

Instead of removing the irrelevant variables, one can reduce the dimensionality of the input space. An effective procedure for performing this operation is *principal components analysis* (PCA). PCA has many advantages and has been used in many applications with great success (Khayamian et al., 2005). This technique has three effects on the data: It orthogonalizes the components of the input vectors (so that they are uncorrelated with each other), it orders the resulting orthogonal components (principal components) so that those with the largest variation come first, and it eliminates those components that contribute the least to variation in the data set. PCA is based on the following assumptions:

- The dimensionality of data can be reduced efficiently by linear transformation.
- Most information is contained in those directions where input data variance is maximum.

The PCA method generates a new set of variables, called *principal components*. Each principal component is a linear combination of the original variables. All the principal components are orthogonal to each other, so there is no redundant information. The principal components as a whole form an orthogonal basis for the space of the data. This approach will result in a significantly reduced set of uncorrelated variables, which will help to reduce the complexity of the network and prevent overfitting (Samarasinghe, 2006).

In applications where wavelet networks are used for the prediction of future values of a target variable, PCA can be proved very useful. On the other hand, in applications where wavelet networks are used for function approximation or sensitivity analysis, PCA can be proved to be cumbersome. The main disadvantage of PCA is that the principal components are usually a combination of all the available variables. Hence, it is often very difficult to distinguish which variable is important and which is not statistically significant. In addition, extra care must be taken when linking to the original variables the information resulting from principal components.

PCA cannot always be used since a linear transformation among the explanatory variables is not always able to reduce the dimension of the data set. Another disadvantage of PCA is the fact that the directions maximizing variance do not always maximize information. Finally, PCA is an orthogonal linear transformation, whereas the use of a wavelet network implies a nonlinear relationship between the explanatory variables and the dependent variable.

Wei et al. (2004) present a novel approach to term and variable selection. This method applies locally linear models together with orthogonal least squares to determine which of the input variables are significant. This algorithm ranks the variables and determines the amount of a system's output variance that can be explained by each term. The method assumes that nonlinearities in the system are relatively smooth. Then, local linear models are fitted in each interval. However, the number of locally linear models increases exponentially as the number of intervals for each independent variable is increased (Wei et al., 2004). Also, selecting the optimal operating regions for the locally piecewise linear models is usually computationally expensive (Wei et al., 2004).

A similar approach was presented by Wei and Billings (2007) based on feature subset selection. In feature selection an optimal or suboptimal subset of the original features is selected (Mitra et al., 2002). More precisely, Wei and Billings (2007) presented a forward orthogonal search algorithm by maximizing the overall dependency to detect the significant variables. This algorithm can produce efficient subsets with a direct link back to the underlying system. The method proposed assumes a linear relationship between sample features. However, this assumption is not always true, and the method may lead to a wider subset of explanatory variables.

SENSITIVITY CRITERIA

Alternatively, one can quantify the average effect of each input variable, x_j, on the output variable, y. The sensitivity of the wavelet network output according to small input perturbations of variable x_j can be estimated either by applying the *average derivative* (AvgD) or the *average elasticity* (AvgL), where the effect is presented as a percentage and is given by the following equations:

$$\mathrm{AvgD}(x_j) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \hat{y}}{\partial x_{ij}} \tag{5.1}$$

$$\mathrm{AvgL}(x_j) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \hat{y}}{\partial x_{ij}}\frac{x_{ij}}{\hat{y}} \tag{5.2}$$

Note that for an estimation of the average elasticity it is assumed that $\hat{y}_i \neq 0$. The average elasticity, apart from having a natural interpretation, takes into account the differences in magnitude between y and x . For example, assume that for the pairs $z_1 = \{x_1, y_1\}$ and $z_2 = \{x_2, y_2\}$ the derivatives of $\hat{y}$ with respect to x are equal. Then, in general, the same change in x (in value or percentage change) will not induce the same change in y, because of the differences in magnitude between x_1 and x_2 and $y_1$and y_2, even if $dy_1/dx_1 = dy_2/dx_2$.

Although the average elasticity conveys more information, in both criteria cancellations between negative and positive values are possible. Natural extensions of the criteria above are the *average derivative magnitude* (AvgDM) and the *average elasticity magnitude* (AvgLM):

$$\mathrm{AvgDM}(x_j) = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\partial \hat{y}}{\partial x_{ij}}\right| \tag{5.3}$$

$$\mathrm{AvgLM}(x_j) = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\partial \hat{y}}{\partial x_{ij}}\right|\left|\frac{x_{ij}}{\hat{y}}\right| \tag{5.4}$$

Note that for estimation of the average elasticity magnitude, it is assumed that $\hat{y}_i \neq 0$. Equations (5.1) to (5.4) utilize the average derivative of the output of the wavelet network with respect to each explanatory variable. As in averaging procedures, a lot of information is lost, so additional criteria are introduced.

The *maximum* and *minimum derivative* (MaxD, MinD) or the *maximum* and *minimum derivative magnitude* (MaxDM, MinDM) give additional insight into the sensitivity of the wavelet network output to each explanatory variable. However, these criteria generally cannot be used on their own since they are appropriate only for some applications and are sensitive to inflection points (Zapranis and Refenes, 1999).

$$\mathrm{MaxD}(x_j) = \max_{i=1\ldots n}\left\{\frac{\partial \hat{y}}{\partial x_{ij}}\right\} \tag{5.5}$$

$$\mathrm{MinD}(x_j) = \min_{i=1\ldots n}\left\{\frac{\partial \hat{y}}{\partial x_{ij}}\right\} \tag{5.6}$$

$$\mathrm{MaxDM}(x_j) = \max_{i=1\ldots n}\left\{\left|\frac{\partial \hat{y}}{\partial x_{ij}}\right|\right\} \tag{5.7}$$

$$\mathrm{MinDM}(x_j) = \min_{i=1\ldots n}\left\{\left|\frac{\partial \hat{y}}{\partial x_{ij}}\right|\right\} \tag{5.8}$$

A way of “standardizing” $\partial \hat{y}\big/\partial x_{ij}$ is to compute the x_j’s *average contribution to the magnitude of the gradient vector* since the local gradient is calculated from all derivatives (Zapranis and Refenes, 1999). Locally, the relative contribution of $\partial \hat{y}\big/\partial x_{ij}$ to the gradient magnitude is given by the ratio

$$r_{ij} = \frac{(\partial \hat{y}_i/\partial x_{ij})^2}{\|\nabla \hat{y}_i\|^2} = \frac{(\partial \hat{y}_i/\partial x_{ij})^2}{\sum_{i=1}^{m}(\partial \hat{y}_i/\partial x_{ij})^2} \tag{5.9}$$

and the x_j’s average contribution to the gradient magnitude by

$$\mathrm{AvgSTD}(x_j) = \frac{1}{n}\sum_{i=1}^{n} r_{ij} \tag{5.10}$$

Alternatively, the *standard deviation of the derivatives* across the sample measures the dispersion of the derivatives around their mean and is given by

$$\mathrm{SDD}(x_j) = \left[\frac{1}{n}\sum_{j=1}^{n}\left(\frac{\partial \hat{y}}{\partial x_{ij}} - \mathrm{AvgD}(x_j)\right)^2\right]^{1/2} \tag{5.11}$$

Finally, the standard deviation per unit of sensitivity, called the *coefficient of variation*, is given by

$$\mathrm{CVD}(x_j) = \frac{\mathrm{SDD}(x_j)}{\mathrm{AvgD}(x_j)} \tag{5.12}$$

MODEL FITNESS CRITERIA

As an alternative to sensitivity criteria, model fitness criteria such as the *sensitivity-based pruning* (SBP) proposed by Moody and Utans (1992) or the effect on the *coefficient of determination* of a small pertubation of x can be used.

Sensitivity-Based Pruning The SBP method quantifies a variable's relevance to the model by the effect on the empirical loss of the replacement of that variable by its mean. The SBP is given by

$$\mathrm{SBP}(x_j) = L_n(\mathbf{x}; \hat{\mathbf{w}}_n) - L_n(\bar{\mathbf{x}}^{(j)}; \hat{\mathbf{w}}_n) \tag{5.13}$$

where

$$\bar{\mathbf{x}}^{(j)} = (x_{1,t}, x_{2,t}, \ldots, \bar{x}_j, \ldots, x_{m,t}) \tag{5.14}$$

and

$$\bar{x}_j = \frac{1}{n}\sum_{t=1}^{n} x_{j,t} \tag{5.15}$$

Additional criteria can be used, such as those presented by Dimopoulos et al. (1995).

(Adjusted) Coefficient of Determination A model-fitness sensitivity criterion that is simpler to interpret is the effect of a variable on the sample coefficient of determination R^2, as measured by the derivative of R^2 with respect to x_j. The coefficient of determination is defined as the ratio

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\mathrm{SSR}}{\mathrm{SST}} \tag{5.16}$$

where SSR and SST are the regression sum of squares and the total sum of squares, respectively, given by

$$\mathrm{SSR} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \tag{5.17}$$

$$\mathrm{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{5.18}$$

The coefficient of determination can also be written as

$$R^2 = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}} \tag{5.19}$$

where SSE is the sum of squared residuals, given by

$$\mathrm{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{5.20}$$

Note that SST is fixed for a given sample; hence, the derivative $\partial R^2/\partial x_j$ can be computed by the following relationship:

$$\frac{\partial R^2}{\partial x_j} = \frac{1}{\mathrm{SST}} \frac{\partial \mathrm{SSR}}{\partial x_j} \tag{5.21}$$

where

$$\frac{\partial \mathrm{SSR}}{\partial x_j} = 2 \sum_{i=1}^{n} \frac{\partial \hat{y}_i}{\partial x_j} (\hat{y}_i - \bar{y}) \tag{5.22}$$

Hence, relationship (5.21) can be written as

$$\frac{\partial R^2}{\partial x_j} = \frac{2}{\mathrm{SST}} \sum_{i=1}^{n} \frac{\partial \hat{y}_i}{\partial x_j} (\hat{y}_i - \bar{y}) \tag{5.23}$$

Alternatively, the adjusted coefficient of determination, $\bar{R}^2$, can be estimated. The adjusted coefficient of determination is an attempt to take into account automatically the phenomenon of R^2 and increasing spuriously when extra explanatory variables are added to the model. $\bar{R}^2$ is given by

$$\bar{R}^2 = 1 - \frac{\mathrm{SSE}/\mathrm{df}_e}{\mathrm{SST}/\mathrm{df}_t} = 1 - \frac{\mathrm{SSE}/(n-p)}{\mathrm{SST}/(n-1)} \tag{5.24}$$

where df is the number of degrees of freedom, n is the number of training patterns, and p is the number of parameters that are adjusted during the training phase of the wavelet network. The number of parameters is computed by

$$p = 1 + 2\lambda j + \lambda + j = 2(\lambda + j) + j + 1 \tag{5.25}$$

where λ is the number of hidden units and j is the number of input variables.

ALGORITHM FOR SELECTING THE SIGNIFICANT VARIABLES

To test statistically whether or not a variable is insignificant and can be removed from the training data set, the distributions of the criteria presented in the preceding section are needed. Without the distribution of the preferred measure of relevance, it is not clear if the effects of the variable x_i on y are statistically significant (Zapranis and Refenes, 1999). More precisely, the only information obtained by criteria described in the preceding section is how sensitive the dependent variable is to small perturbations of the independent variable. It is clear that the smaller the value of the preferred criterion, the less significant the corresponding variable is. However, there is no information as to whether or not this variable should be removed from the model.

We use the bootstrap method to approximate asymptotically the distribution of the measures of relevance. More precisely, a number of bootstrapped training samples can be created by the original training data set. The idea is to estimate the preferred criterion on each bootstrapped sample. If the number of the bootstrapped samples is large, a good approximation of the empirical distribution of the criterion is expected. Obtaining an approximation of the empirical distributions, confidence intervals and hypothesis tests can be constructed for the value of the criterion. The variable selection algorithm is explained analytically below and is illustrated in Figure 5.1.

The algorithm starts with a training sample that consists of all available explanatory variables.

1. Create B bootstrapped training samples from the original data set.
2. Identify the correct topology of the wavelet network following the procedure described Chapter 4 and estimate the prediction risk.
3. Estimate the preferred measure of relevance for each explanatory variable for each of the B bootstrapped training samples.
4. Calculate the p-values of the measure of relevance.
5. Test if any explanatory variables have a p-value greater than 0.1. If variables with a p-value greater than 0.1 exist, the variable with the largest p-value is removed from the training data set; else, the algorithm stops.
6. Estimate the prediction risk and the new p-values of the reduced model. If the new estimated prediction risk is smaller than the prediction risk multiplied by a

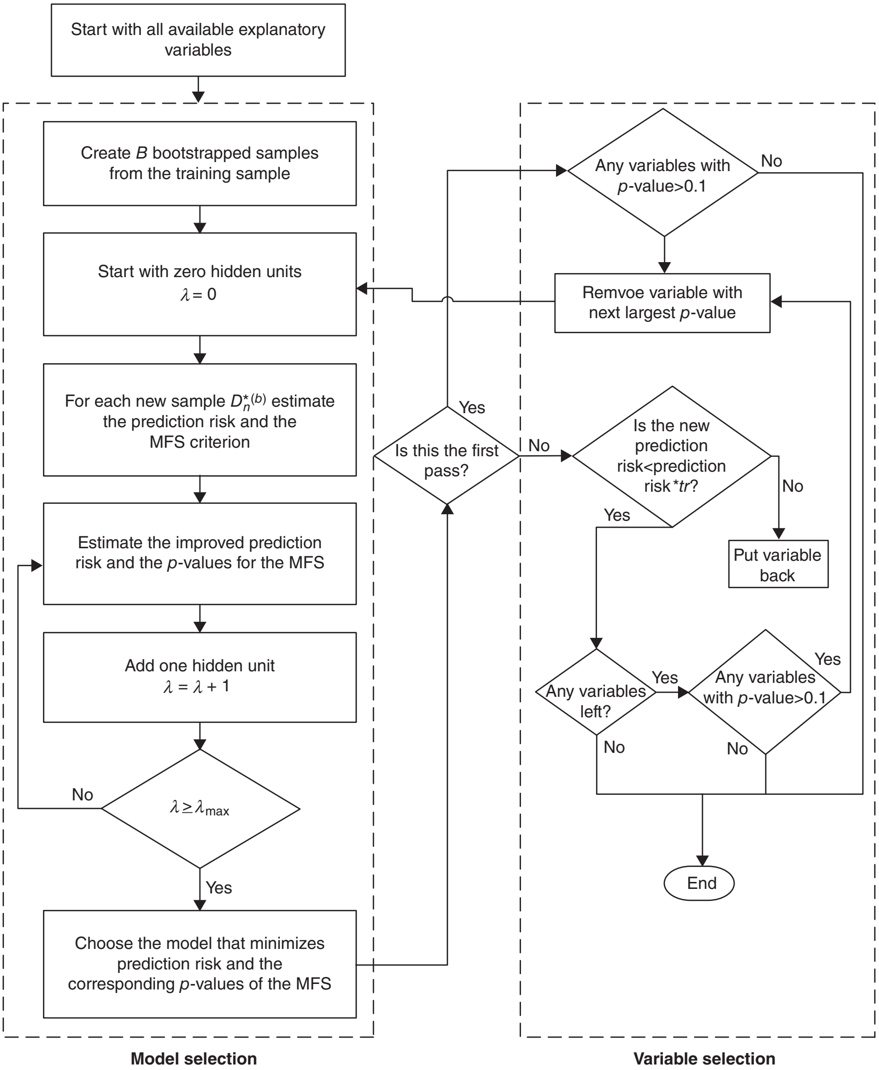
Start with all available explanatory variables
Create B bootstrapped samples from the training sample
Start with zero hidden units $\lambda = 0$
For each new sample $D_n^{*(b)}$ estimate the prediction risk and the MFS criterion
Estimate the improved prediction risk and the p-values for the MFS
Add one hidden unit $\lambda = \lambda + 1$
$\lambda \geq \lambda_{\max}$
No
Yes
Choose the model that minimizes prediction risk and the corresponding p-values of the MFS
Is this the first pass?
Yes
No
Any variables with p-value>0.1
No
Remvoe variable with next largest p-value
Is the new prediction risk<prediction risk*tr?
Yes
No
Put variable back
Any variables left?
Yes
No
Any variables with p-value>0.1
Yes
No
End
Model selection
Variable selection

Figure 5.1 *Model identification: model selection and variable selection algorithms.*

threshold (usually, 5%), the decision of removing the variable was correct and we return to step 5.

7. If the new prediction risk is greater than the new prediction risk multiplied by a threshold (usually, 5%), the decision to remove the variable was wrong and the variable must be reintroduced to the model. In this case the variable with the next largest p-value which is also greater than 0.1 is removed from the training sample and we return to step 6. If the remaining variables have p-values smaller than 0.1, the algorithm stops.

Resampling Methods for the Estimation of Empirical Distributions

To have a good estimation of the prediction risk as well as an approximation of the distribution of the measure of relevance, a large number of bootstrapped samples B are needed. As B increases, the accuracy of the algorithm also increases, but so does the computational burden. Zapranis and Refenes (1999) presented two different bootstrap methods, *local bootstrap* and *parametric sampling*, that are significantly less computationally expensive.

Since a unique strong solution of the loss function $L(\mathbf{x}; \hat{\mathbf{w}}_n)$ does not exist, the local bootstrap is proposed by Zapranis and Refenes (1999). In local bootstrapping, first a network is trained where the original training sample is used as an input and the vector $\hat{\mathbf{w}}_n$ that minimizes the loss function is estimated. Then new samples are generated from the original training patterns using the bootstrap method. To train the new samples, the neural networks are not initialized randomly; on the contrary, the initial conditions are given by the vector $\hat{\mathbf{w}}_n$ estimated by the initial training sample. Starting the training of the neural network very close to $\hat{\mathbf{w}}_n$, the probability of the convergence of the neural network to another local minimum is reduced significantly.

A novel approach (parametric sampling) is also presented by Zapranis and Refenes (1999). The distribution of the weights of a neural network is known. As has been shown by Galland and White (1988) and White (1989), the asymptotic distribution of $\sqrt{n}(\hat{\mathbf{w}}_n - \mathbf{w}_0)$ is a multivariate normal distribution with zero mean and known covariance matrix $\mathbf{C}$, where $\hat{\mathbf{w}}_n$ is the estimated vector and $\mathbf{w}_0$ is the true vector of parameters that minimizes the loss function. Since $\mathbf{w}_0$ is not known, an estimator $\hat{\mathbf{C}}_n$ of the covariance matrix has to be used.

A basic requirement of parametric sampling is *locally identified models*, models without superfluous connections. The most important assumption is that the network was not converged in a flat minimum (i.e., $\hat{\mathbf{w}}_n$ has to be a local unique solution). This can be avoided if the irrelevant connections are removed using pruning techniques. As a result, an estimate of the standard error of any function of $\hat{\mathbf{w}}_n$ can be estimated robustly (Zapranis and Refenes, 1999).

In the framework above, instead of creating new bootstrapped training samples, Zapranis and Refenes (1999) propose sampling from the distribution of $\hat{\mathbf{w}}_n$ (parametric sampling). As a result, a very large number of parameter vectors $\hat{\mathbf{w}}_n^{(a)}$can be created. Then any function of $\hat{\mathbf{w}}_n$ can be estimated using the bootstrapped parameter vectors $\hat{\mathbf{w}}_n^{(a)}$. This procedure can be applied to any function, like that of the measures of significance of the explanatory variables presented in the preceding section. The proposed scheme is orders of magnitude faster than alternative methods since there is no need to train new networks. For each new parameter vector the corresponding model fitness or sensitivity criterion is evaluated. However, parametric sampling is computationally more complex, since the computation and inversion of the Hessian matrix of the loss function must be estimated in order to compute $\hat{\mathbf{C}}_n$.

However, the bootstrapped samples may differ significantly from the original sample. Hence, applying local bootstrapping or parametric sampling may lead to wavelets outside their effective support (i.e., wavelets with value of zero), since wavelets are local functions with limited duration. In addition, in contrast to the case of neural networks, the asymptotic distribution of the weights of a wavelet network is not known. These observations constitute both local bootstrapping and parametric sampling inappropriate for wavelet networks.

Alternatively, new samples from training patterns can be constructed. This can be done by applying bootstrapping from pairs and training a wavelet network for each sample. Since the initialization of a wavelet network is very good, this procedure is not prohibitively expensive. The computational cost of this algorithm is B, the number of wavelet networks that must be trained. As *v-fold* cross-validation is much less computationally expensive and performs as well as bootstrapping, an approach where 50 new training samples are created according to v-fold cross-validation can also be employed.

EVALUATING THE VARIABLE SIGNIFICANCE CRITERIA

In this section the algorithm proposed in the preceding section to select the significant explanatory variables is evaluated. More precisely, the eight sensitivity criteria and the model fitness sensitivity criterion are evaluated for the two functions, $f(x)$ and $g(x)$, presented in Chapter 3.

Case 1: Sinusoid and Noise with Decreasing Variance

First, a second variable is created, which was drawn randomly from the uniform distribution within the range (0,1). Both variables are considered significant and constitute the training patterns $(\mathbf{x}_i, y_i)$ of the training data set, where $\mathbf{x}_i = \{x_{1,i}, x_{2,i}\}$and y_i are the target values. A wavelet network is trained in order to learn the target function $f(x)$, where both x_1 and x_2 are introduced to the wavelet network as input patterns. The BE algorithm was used for initialization of the wavelet network. Using cross-validation and bootstrapping, the prediction risk is minimized when 3 hidden units are used and is 0.04194. The network converges after 3502 iterations. Comparing the results with the findings in the preceding section, it is clear that including an irrelevant variable in our model increases the model complexity and the training time while reducing the predictive power of the model. Note that as presented in Chapter 4, when only the relevant variable was used, the optimal number of hidden units was 2, while the network converged after only one iteration.

After the wavelet network is fully trained, the various measures of relevance presented in the preceding section can be estimated. The weights of the direct connections between the explanatory variables and the network output are presented in Table 5.1. In addition, the following sensitivity criteria are also presented: MaxD, the MinD, MaxDM, MinDM, AvgD, AvgDM, AvgL, and AvgLM. Finally, the SBP criterion can be found in the final column of the table. The first part of the table refers to the full model, where both variables x_1 and x_2 were used, while the second part refers to the reduced model, where only the significant variable x_1 was used.

TABLE 5.1 Sensitivity Measures for the First Case[a]

	$w_i^{[0]}$	MaxD	MinD	MaxDM	MinDM	AvgD	AvgDM	AvgL	AvgLM	SBP
Full model (two variables)										
X_1	0.0161	1.3962	−1.3459	1.3962	0.0005	−0.0529	0.6739	0.2127	1.6323	0.0953
X_2	0.0186	0.4964	−0.7590	0.7590	0.0002	0.0256	0.0915	0.0781	0.1953	0.0001
Reduced model (one variable)										
X_1	0.1296	1.1646	−1.1622	1.1644	0.0014	0.0841	0.7686	0.3165	1.3510	0.0970

[a] $w_i^{[0]}$, the linear connection between the input and output variables; MaxD, maximum derivative; MinD, minimum derivative; MaxDM, maximum derivative magnitude; MinDM, minimum derivative magnitude; AvgD, average derivative; AvgDM, average derivative magnitude; AvgL, average elasticity; AvgLM, average elasticity magnitude; SBP, sensitivity-based pruning.

Examining the direct connections of the weights between the explanatory variables and the network output, $w_i^{[0]}$, we conclude that both variables have the same significance, since both weights have almost the same value, with $w_2^{[0]}$ being slightly larger. Closer inspection of Table 5.1 reveals a contradiction between AvgL and AvdD. The AvgL value for the first variable is 0.2127 and for the second variable is 0.0781, indicating that x_1 is more significant than x_2. On the other hand, AvgD for x_2 is −0.0529 and for x_2 is 0.0256, indicating that changes in x_1 have an opposite effect on the dependent variable. In these two criteria, cancellations between negative and positive values are possible, so it is essential also to examine the AvgLM and AvgDM. Both criteria are significantly larger in the case of the first variable. The same results are obtained for the remaining sensitivity criteria. Note that MinD indicates as significant the variable with the smaller value. Finally, the SBP for the first variable is 0.0953 and for the second variable is only 0.0001, indicating that the second variable has a negligible effect on the performance of the wavelet network. Next, observing the values of the various criteria for the reduced model in Table 5.1, we conclude that the value of the weight, MinDM, and AvgD are affected by the presence of the second variable.

The previous simple case indicates that of the 10 criteria listed, only three produced robust results: AvgDM, AvgLM, and SBP. However, perhaps with the exception of SBP, it is unclear if the second variable should be removed from the training data set.

Next, the algorithm described in the preceding section will be utilized to estimate the p-values of each criterion. More precisely, the bootstrap and cross-validation methods will be applied to estimate the asymptotic distributions of the various criteria presented in Table 5.1. The mean, standard deviation, and p-values for all sensitivity and model fitness measures for the two variables of the first case are presented in Table 5.2. Using cross-validation, 50 new samples were created to approximate the empirical distributions of the various criteria, and the corresponding criteria and p-values were calculated. As expected, the average values of the criteria are similar to those presented in Table 5.1. Observing Table 5.2, it is clear that x_1

TABLE 5.2 Variable Significance Testing for the First Case Using Cross-Validation[a]

	MaxD	MinD	MaxDM	MinDM	AvgD	AvgDM	AvgL	AvgLM	SBP
Full model (two variables)									
X_1	0.9947	−1.5589	1.5589	0.0037	−0.1369	0.6898	−0.0753	1.2667	0.0967
Std.	0.0332	0.0120	0.0120	0.0023	0.0091	0.0051	0.0258	0.0375	0.0006
p-Value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
X_2	0.6231	−0.6099	0.6231	0.0001	0.0575	0.1253	0.0976	0.2137	−0.0001
Std.	0.0342	0.0182	0.0170	0.0001	0.0166	0.0031	0.0235	0.0078	0.0001
p-Value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0153	0.0000	0.6019
Reduced model (one variable)									
X_1	—	—	—	—	—	—	—	—	0.0970
Std.	—	—	—	—	—	—	—	—	0.0006
p-Value	—	—	—	—	—	—	—	—	0.0000

[a] MaxD, maximum derivative; MinD, minimum derivative; MaxDM, maximum derivative magnitude; MinDM, minimum derivative magnitude; AvgD, average derivative; AvgDM, average derivative magnitude; AvgL, average elasticity; AvgLM, average elasticity magnitude; SBP, sensitivity-based pruning.

has a larger impact on output y. However, all eight sensitivity measures consider both variables as significant predictors. As discussed previously, these criteria are application dependent; model fitness criteria are much better suited for testing the significance of the explanatory variables (Zapranis and Refenes, 1999). Indeed, the p-value for x_2 using SBP is 0.6019, indicating that this variable must be removed from the model. On the other hand, the p-value for x_1 using SBP is zero, indicating that x_1 is very significant. Finally, the p-value of x_1 using SBP in the reduced model is zero, indicating that x_1 is still very significant. Moreover, the average value of SBP is almost the same in the full and reduced models.

Next, the bootstrap method will be need to estimate the asymptotic distributions of the various criteria. To approximate the empirical distributions of the various criteria, 50 new bootstrapped samples were created, and their corresponding p-values are presented in Table 5.3. In the table the same analysis is repeated for the first case, but the random samples were created using bootstrapping. As in the cross-validation approach, 50 new bootstrapped samples were created to approximate the empirical distributions of the various criteria. A closer inspection of Table 5.3 reveals that MaxD, MinD, MaxDM, MinDM, AvgDM, and AvgLM suggest that both variables are significant and must remain in the model. On the other hand, the p-values obtained using the AvgL criterion wrongly suggests that the variable x_1 must be removed from the model. Finally, SBP and AvgD suggest correctly that x_2 must be removed from the model. More precisely, the p-values obtained using AvgD are 0.0614 and 0.3158 for x_1 and x_2, respectively, while the p-values obtained using SBP are 0 and 0.9434 for x_1 and x_2, respectively. Finally, the p-value of x_1 using SBP in the reduced model is zero, indicating that x_1 is still very significant. However, while the average value

TABLE 5.3 Variable Significance Testing for the First Case Using Bootstrapping[a]

	MaxD	MinD	MaxDM	MinDM	AvgD	AvgDM	AvgL	AvgLM	SBP
Full model (two variables)									
X_1	1.6242	−2.1524	2.2707	0.0031	−0.1079	0.6998	−0.0267	1.3498	0.0982
Std.	1.3929	2.3538	2.4426	0.0029	0.0758	0.0391	0.1651	0.4161	0.0045
p-Value	0.0000	0.0000	0.0000	0.0000	0.0614	0.0000	**0.6039**	0.0000	0.0000
X_2	1.1038	−1.2013	1.4472	0.0003	0.0402	0.1369	0.1033	0.2488	0.0011
Std.	1.4173	2.6560	2.8320	0.0003	0.0477	0.0277	0.1010	0.1244	0.0013
p-Value	0.0000	0.0000	0.0000	0.0179	0.3158	0.0000	0.4610	0.0000	0.9434
Reduced model (one variable)									
X_1	—	—	—	—	0.0800	—	—	—	0.0988
Std.	—	—	—	—	0.0433	—	—	—	0.0051
p-Value	—	—	—	—	0.0000	—	—	—	0.0000

[a] MaxD, maximum derivative; MinD, minimum derivative; MaxDM, maximum derivative magnitude; MinDM, minimum derivative magnitude; AvgD, average derivative; AvgDM, average derivative magnitude; AvgL, average elasticity; AvgLM, average elasticity magnitude; SBP, sensitivity-based pruning.

of SBP is almost the same in the full and reduced models, the average value of AvgD is completely different in magnitude and sign.

The correctness of removing a variable from the model should always be tested further. As discussed in the preceding section, this can be done either by estimating the prediction risk or the $\bar{R}^2$of the reduced model. The prediction risk in the reduced model was reduced to 0.0396, while it was 0.0419 in the full model. Moreover, the $\bar{R}^2$ increased to 70.8% in the reduced model, while it was 69.8% in the full model. The results indicate that the decision to remove x_2 was correct.

Case 2: Sum of Sinusoids and Cauchy Noise

The same procedure is repeated for the second case, where a wavelet network is used to learn the function $g(x)$ from noisy data. First, a second variable is created which was drawn randomly from the uniform distribution within the range (0,1). Both variables are considered significant and constitute the training patterns. A wavelet network is trained to learn the target function $g(x)$, where both x_1 and x_2 are introduced to the wavelet network as inputs patterns. The BE algorithm was used for initialization of the wavelet network. Using cross-validation and bootstrapping, the prediction risk is minimized when 10 hidden units are used and is 0.00336. The network approximation converges after 18,811 iterations. Again the inclusion of an irrelevant variable to our model increased the model complexity and the training time while reducing the predictive power of the model. Note that when only the relevant variable was used for the training of the wavelet network, only 8 hidden units were used, whereas the wavelet network was converged after only 1107 iterations.

TABLE 5.4 Sensitivity Measures for the Second Case[a]

	$w_i^{[0]}$	MaxD	MinD	MaxDM	MinDM	AvgD	AvgDM	AvgL	AvgLM	SBP
Full model (two variables)										
X_1	−0.0001	1.4991	−1.4965	1.4991	0.0001	0.0032	0.5517	−1.1417	6.5997	0.4202
X_2	0.0124	3.4623	−2.5508	3.4623	0.0001	0.0261	0.2691	−0.0095	0.1898	0.0002
Reduced model (one variable)										
X_1	−0.0963	1.6802	−1.5662	1.6801	0.0019	0.0011	0.6031	−0.8662	9.7935	0.4707

[a] $w_i^{[0]}$, the lineral connection between the input and output variables; MaxD, maximum derivative; MinD, minimum derivative; MaxDM, maximum derivative magnitude; MinDM, minimum derivative magnitude; AvgD, average derivative; AvgDM, average derivative magnitude; AvgL, average elasticity; AvgLM, average elasticity magnitude; SBP, sensitivity-based pruning.

Table 5.4 presents the weights of the direct connections between the explanatory variables and the network output, MaxD, MinD, MaxDM, MinDM, AvgD, AvgDM, AvgL, and AvgLM sensitivity criteria, and the SBP model fitness criterion. The first part of the table refers to the full model, where both variables x_1 and x_2 were used; the second part refers to the reduced model, where only the variable x_1 is used.

From Table 5.4 it is clear that the value of the weight of the direct connections between the first variable and the network output is smaller than the weight of the second variable. A closer inspection of the table reveals that almost all measures of relevance wrongly identify the second variable as more significant than x_1. The only exceptions are the AvgDM and AvgLM criteria. Both criteria give a significantly larger value in x_1. Finally, SBP for the first variable is 0.4202 for the first variable, while for the second variable it is only 0.0002, indicating that the second variable has a negligible effect on the performance of the wavelet network. Next, observing the values of the various criteria for the reduced model on the table, we conclude that the value of the weight, MinDM, and AvgD are affected again by the presence of the second variable.

Next, we estimate the p-values of the various criteria for the second case. The standard deviation and p-values for all sensitivity and model fitness measures for the two variables of the second case are presented in Table 5.5. Using cross-validation, 50 new samples were created to approximate the empirical distributions of the various criteria, and the corresponding criteria and p-values were calculated. The table shows that only AvgLM and SBP indentified the insignificant variable correctly. The p-values are 0.4838 and 0.6227, respectively, for the two criteria for x_2, while in the reduced model the p-values of x_1 are zero. On the other hand, MinD and AvgL wrongly suggested that x_1 should be removed from the model. Finally, the remaining criteria, MaxD, MinD, MaxDM, AvgD, and AvgDM, suggest that both variables are significant and should remain in the model.

In Table 5.6 the same analysis is repeated for the second case, but the random samples were created using bootstrapping. As in the CV approach, 50 new bootstrapped

TABLE 5.5 Variable Significance Testing for the Second Case Using Cross-Validation[a]

	MaxD	MinD	MaxDM	MinDM	AvgD	AvgDM	AvgL	AvgLM	SBP
Full model (two variables)									
X_1	1.5624	−1.8437	1.8467	0.0016	−0.0192	0.5865	−2.7769	17.5475	0.4558
Std.	0.0061	0.1623	0.1561	0.0009	0.0038	0.0021	17.5304	14.7331	0.0035
p-Value	0.0000	0.0000	0.0000	**0.6897**	0.0000	0.0000	**0.7184**	0.1258	0.0000
X_2	0.7745	−1.5054	1.5054	0.0004	−0.1797	0.2349	0.0091	0.2438	0.0002
Std.	0.0214	0.0843	0.0843	0.0003	0.0056	0.0038	0.0659	0.0586	0.0002
p-Value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.3539	0.4838	0.6227
Reduced model (one variable)									
X_1	—	—	—	—	—	—	—	9.4370	0.4776
Std.	—	—	—	—	—	—	—	1.8363	0.0075
p-Value	—	—	—	—	—	—	—	0.0000	0.0000

[a] MaxD, maximum derivative; MinD, minimum derivative; MaxDM, maximum derivative magnitude; MinDM, minimum derivative magnitude; AvgD, average derivative; AvgDM, average derivative magnitude; AvgL, average elasticity; AvgLM, average elasticity magnitude; SBP, sensitivity-based pruning.

samples were created to approximate the empirical distributions of the various criteria. In the table the analysis for the second case is presented. A closer inspection of the table reveals that MaxD, MinD, AvgDM, and AvgLM suggest that both variables are significant and must remain in the model. On the other hand, the p-values obtained using the AvgL and AvgD criteria wrongly suggest that the variable x_1 must be

TABLE 5.6 Variable Significance Testing for the Second Case Using Bootstrapping[a]

	MaxD	MinD	MaxDM	MinDM	AvgD	AvgDM	AvgL	AvgLM	SBP
Full model (two variables)									
X_1	1.6485	−1.8391	1.9459	0.0006	0.0225	0.5412	0.2908	8.9262	0.4191
Std.	0.3555	0.7505	0.7475	0.0008	0.0736	0.0524	7.0110	5.9525	0.0589
p-Value	0.0000	0.0000	0.0000	0.2867	**0.9877**	0.0000	**0.8708**	0.0000	0.0000
X_2	10.0490	−7.7106	11.4443	0.0007	0.0269	0.4564	−0.1217	0.6045	0.0024
Std.	16.2599	9.5366	16.9065	0.0005	0.0923	0.2912	0.5508	0.7338	0.0085
p-Value	0.07838	0.0762	0.1597	0.4158	0.6686	0.0000	0.7864	0.0000	0.8433
Reduced model (one variable)									
X_1	—	—	1.7261	0.0009	—	—	—	—	0.4779
Std.	—	—	0.0916	0.0008	—	—	—	—	0.0255
p-Value	—	—	0.0000	0.1795	—	—	—	—	0.0000

[a] MaxD, maximum derivative; MinD, minimum derivative; MaxDM, maximum derivative magnitude; MinDM, minimum derivative magnitude AvgD, average derivative; AvgDM, average derivative magnitude; AvgL, average elasticity; AvgLM, average elasticity magnitude; SBP, sensitivity-based pruning.

removed from the model. Finally, SBP, MaxD, and Min DM suggest correctly that x_2 is not a significant variable and can be removed from the model. More precisely, the p-values obtained using MaxDM are 0 and 0.1597 for x_1 and x_2, respectively, while the p-values obtained using MinDM are 0.2867 and 0.4158 for x_1 and x_2, respectively. Finally, the p-values obtained using SBP are 0 and 0.8433 for x_1 and x_2, respectively. Examining the reduced model, where only x_1 is used for the training of the WN, the p-values are zero for x_1 when the MaxDM or SBP criteria are used. On the other hand, the p-value for x_1 is 0.1795 when the MinDM is used, indicating that x_1 is insignificant and should also be removed from the model.

Next, the correctness of removing a variable from the model is tested further. As discussed in the preceding section, this can be done either by estimating the prediction risk or the $\bar{R}^2$of the reduced model. The prediction risk in the reduced model was reduced to 0.0008, while it was 0.0033 in the full model. Moreover, the $\bar{R}^2$ increased to 99.7% in the reduced model, while it was 99.2% in the full model.

CONCLUSIONS

The criteria presented in the preceding section introduce a measure of relevance between the input and output variables. These criteria can be used for data preprocessing, sensitivity analysis, and variable selection. In this chapter we developed an algorithm for variable selection based on the empirical distribution of these criteria and examined their performance.

The results from the previous simulated experiments indicate that SBP gives constantly correct and robust results. In every case, the SBP criterion correctly identified the irrelevant variable. Moreover, the SBP criterion was stable and had the same magnitude and sign in both the full and reduced models.

The results of the previous cases indicate that when our algorithm is employed and the p-values are estimated, the performance of the remaining sensitivity criteria is unstable. In general, the sensitivity criteria were not able to identify the insignificant variable. Moreover, they often suggested removal of the significant variable x_1. The sensitivity criteria are application dependent, and extra care must be taken when they are used (Zapranis and Refenes, 1999). As their name suggests, they are more appropriate for use in sensitivity analysis than in variable significance testing.

Finally, when the bootstrap method was used, the standard deviation of each criterion was constantly significantly larger than the values obtained when the cross-validation was used. Bootstrapped samples contain more variability and may differ significantly from the original sample. As a result, an unbiased empirical distribution of the corresponding statistic is obtained.

REFERENCES

Dimopoulos, Y., Bourret, P., and Lek, S. (1995). "Use of some sensitivity criteria for choosing networks with good generalization ability." *Neural Processing Letters*, 2(6), 1–4.

Galland, A. R., and White, H. (1988). *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*. Basil Blackwell, Oxford.

Khayamian, T., Ensafi, A. A., Tabaraki, R., and Esteki, M. (2005). "Principal component-wavelet networks as a new multivariate calibration model." *Analytical Letters*, 38(9), 1447–1489.

Mitra, P., Murthy, C. A., and Pal, S. K. (2002). "Unsupervised feature selection using feature similarity." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3), 301–312.

Moody, J. E., and Utans, J. (1992). "Principled architecture selection for neural networks: application to corporate bond rating prediction." In *Neural Networks in the Capital Markets*, A. P. Refenes, ed. Wiley, New York.

Samarasinghe, S. (2006). *Neural Networks for Applied Sciences and Engineering*. Taylor & Francis, New York.

Wei, H.-L., and Billings, S. A. (2007). "Feature subset selection and ranking for data dimensionality reduction." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1), 162–166.

Wei, H.-L., Billings, S. A., and Liu, J. (2004). "Term and variable selection for non-linear system identification." *International Journal of Control*, 77(1), 86–110.

White, H. (1989). "An additional hidden unit test for neglected nonlineartiy in multilayer feedforward networks." *IJCNN*, Washington, DC, 451–455.

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