

## Chapter 3

# Automatic Relevance Determination in Economic Modeling

**Abstract** This chapter introduces the Bayesian and the evidence frameworks to construct an automatic relevance determination method. These techniques are described in detail, relevant literature reviews are conducted, and their use is justified. The automatic relevance determination technique is then applied to determine the relevance of economic variables that are essential for driving the consumer price index. Conclusions are drawn and are explained within the context of economic sciences.

### 3.1 Introduction

This chapter presents the Bayesian and the evidence frameworks to create an automatic relevance determination technique. These methods are explained in detail, and pertinent literature reviews are conducted. The automatic relevance determination method is then applied to establish the relevance of economic variables that are critical for driving the consumer price index (CPI).

Shutin et al. (2012) introduced incremental reformulated automatic relevance determination and related this to the incremental version of sparse Bayesian learning. The fast marginal likelihood maximization procedure is an incremental method where the objective function is optimized with respect to the parameters of a single component given that the other parameters are fixed. The procedure is then demonstrated to relate to a series of re-weighted convex optimization problems.

Huang et al. (2012) introduced stochastic optimization using an automatic relevance determination (ARD) prior model and applied this for Bayesian compressive sensing. Compressive sensing is a unique data acquisition procedure where the compression is conducted during the sampling process. In condition monitoring systems, original data compression methods such as compressive sensing are required to decrease the cost of signal transfer and storage. Huang et al. (2012) introduced

Bayesian compressive sensing (BCS) for condition monitoring of signals. The results obtained from the improved BCS technique were better than conventional BCS reconstruction procedures.

Böck et al. (2012) used the ARD method for a hub-centered gene network reconstruction procedure and exploited topology of the of gene regulatory networks by using a Bayesian network. The proposed technique was applied to a large publicly available dataset was able to identify several main hub genes.

Jacobs (2012) applied Bayesian support vector regression with an ARD kernel for modeling antenna input characteristics. The results indicated that Bayesian support vector regression was appropriate for highly non-linear modeling tasks. They observed that the Bayesian framework allowed efficient training of the multiple kernel ARD hyper-parameters.

Shutin et al. (2011) proposed fast variational sparse Bayesian learning with ARD for superimposed signals. They showed that a fast version of variational sparse Bayesian learning can be built using stationary points of the variational update factors with non-informative ARD hyper-priors.

Zhang et al. (2010) applied the Gaussian process classification using ARD for synthetic aperture radar target recognition. The method they proposed implemented kernel principal component analysis to identify sample features and applied target recognition using Gaussian process classification with an ARD function. When they compared this technique to the k-Nearest Neighbor clustering method, a Naïve Bayes classifier, and a Support Vector Machine, the proposed technique was found to be able to automatically select an appropriate model and to optimize hyper-parameters.

Lisboa et al. (2009) applied a neural network model and Bayesian regularization with the typical approximation of the evidence to create an automatic relevance determination system. The model was applied to local and distal recurrence of breast cancer.

Mørupa and Hansena (2009) applied automatic relevance determination for multi-way models, while Browne et al. (2008) applied ARD for identifying thalamic regions concerned in schizophrenia. Other successful applications of ARD include in earthquake early warning systems (Oh et al. 2008), feature correlations investigation (Fu and Browne 2008), ranking the variables to determine ischaemic episodes (Smyrnakis and Evans 2007), the estimation of relevant variables in multichannel EEG (Wu et al. 2010) and the estimation of relevant variables in classifying ovarian tumors (Van Calster et al. 2006).

## 3.2 Mathematical Framework

The automatic relevance determination technique is a process applied to evaluate the relevance of each input variable in its ability to predict a particular phenomenon. In this chapter, we apply ARD to determine the relevance of economic variables in predicting the CPI. ARD achieves this task of ranking input variables

by optimizing the hyper-parameters to maximize the evidence in the Bayesian framework (Marwala and Lagazio 2011). As already described by Marwala and Lagazio (2011), Wang and Lu (2006) successfully applied the ARD technique to approximate influential variables in modeling the ozone layer, while Nummenmaa et al. (2007) successfully applied ARD to a dataset where a male subject was presented with uncontaminated tone and checker board reversal stimuli, individually and in combination, in employing a magnetic resonance imaging-based cortical surface model. Ulusoy and Bishop (2006) applied ARD to categorize relevant features for object recognition of 2D images. One part of the ARD as applied in this chapter is neural networks, the topic of the next sub-section. It should be borne in mind that the discourse of ARD is not only limited to neural networks and can, within the context of artificial intelligence, include subjects such as support vector machines, Gaussian mixture models, and many other models.

### ***3.2.1 Neural Networks***

This section gives a summary of neural networks in the context of economic modeling (Leke and Marwala 2005; Lunga and Marwala 2006). A neural network is an information processing procedure that is inspired by the way that biological nervous systems, like the human brain, process information (Marwala and Lagazio 2011). It is a computer-based mechanism that is aimed at modeling the way in which the brain processes a specific function of consideration (Haykin 1999; Marwala and Lagazio 2011).

The neural network technique is a powerful tool that has been successfully used in mechanical engineering (Marwala and Hunt 1999; Vilakazi and Marwala 2007; Marwala 2012), civil engineering (Marwala 2000, 2001), aerospace engineering (Marwala 2003), biomedical engineering (Mohamed et al. 2006; Russell et al. 2008, 2009a, b), finance (Leke and Marwala 2005; Patel and Marwala 2006; Hurwitz and Marwala 2011; Khoza and Marwala 2012), statistics (Marwala 2009), intrusion detection (Alsharafat 2013), telecommunications (Taspinar and Isik 2013) and political science (Marwala and Lagazio 2011).

Meena et al. (2013) applied fuzzy logic and neural networks for gender classification in speech recognition. To train fuzzy logic and neural networks, features that included the pitch of the speech were used and the tests showed good results. Nasir et al. (2013) applied a multilayer perceptron and simplified fuzzy ARTMAP neural networks for the classification of acute leukaemia cells. The cells were classified into lymphoblast, myoblast, and normal cell; to categorize the severity of leukaemia types, and the results obtained gave good classification performance. Shaltaf and Mohammad (2013) applied a hybrid neural network and maximum likelihood based estimation of chirp signal parameters and observed that a hybrid of neural networks and maximum likelihood gradient based optimization gave accurate parameter approximation for large signal to noise ratio.

In this chapter, neural networks are viewed as generalized regression models that can model any data, linear or non-linear. As described by Marwala and Lagazio (2011), a neural network is made up of four main constituents (Haykin 1999; Marwala 2012):

- the processing units  $u_j$ , where each  $u_j$  has a particular activation level  $a_j(t)$  at any given time;
- weighted inter-connections between several processing units. These inter-connections control how the activation of one unit influences the input for another unit;
- an activation rule, which takes input signals at a unit to yield a new output signal; and
- a learning rule that stipulates how to regulate the weights for a given input/output pair (Haykin 1999).

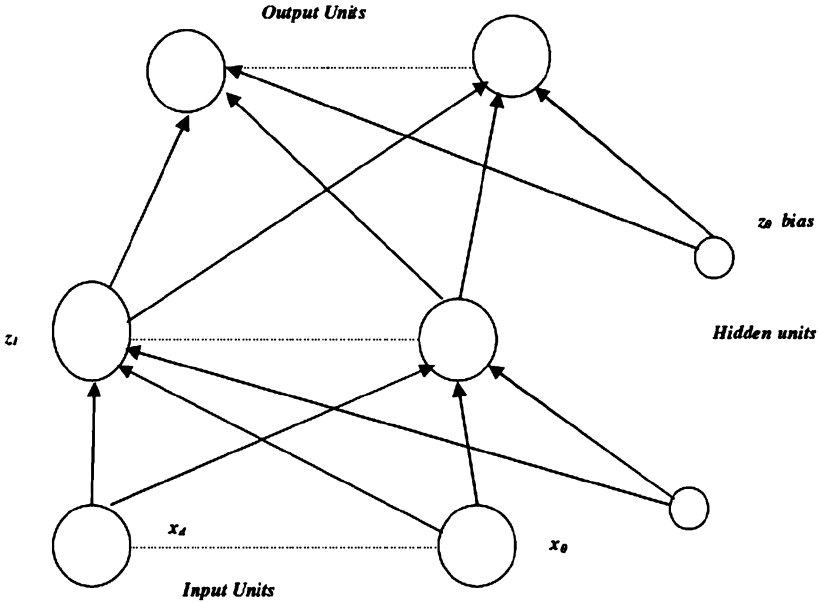
Neural networks are able to derive meaning from complex data and, therefore, can be used to extract patterns and detect trends that are too complex to be detected by many other computer approaches (Hassoun 1995; Marwala 2012). A trained neural network can be viewed as an expert in the type of information it has been set to discover (Yoon and Peterson 1990; Valdés et al. 2012; Sinha et al. 2013). A trained neural network can be used to predict given new circumstances. Neural networks have been applied to model a number of non-linear applications because of their capacity to adapt to non-linear data (Leke et al. 2007; Martínez-Rego et al. 2012).

The topology of neural processing units and their inter-connections can have a deep influence on the processing capabilities of neural networks. Accordingly, there are many different connections that describe how data flows between the input, hidden, and output layers.

There are different kinds of neural network topologies and these include the multi-layer perceptron (MLP) and the radial basis function (RBF) (Bishop 1995; Sanz et al. 2012; Prakash et al. 2012). In this chapter, the MLP was applied to identify the relationship between economic variables and CPI. The motivation for using the MLP was because it offers a distributed representation with respect to the input space due to cross-coupling between hidden units, while the RBF gives only local representation (Bishop 1995; Marwala and Lagazio 2011).

Ikuta et al. (2012) applied the MLP for solving two-spiral problem whereas Rezaeian-Zadeh et al. (2012) applied the MLP and RBF for hourly temperature prediction. Wu (2012) applied a multi-layer perceptron neural network for scattered point data surface reconstruction while Li and Li (2012) implemented the MLP in a hardware application.

The conventional MLP contains hidden units and output units and normally has one hidden layer. The bias parameters in the first layer are presented as mapping weights from an extra input having a fixed value of  $x_0 = 1$ . The bias parameters in the second layer are displayed as weights from an extra hidden unit, with the activation fixed at  $z_0 = 1$ . The model in Fig. 3.1 can take into account the intrinsic dimensionality of the data. Models of this form can approximate any continuous function to arbitrary accuracy if the number of hidden units  $M$  is adequately large.



**Fig. 3.1** Feed-forward multi-layer perceptron network having two layers of adaptive weights (Reprinted with permission from Marwala 2009; Marwala and Lagazio 2011)

The size of the MLP may be increased by permitting for a number of layers, however, it has been verified by the universal approximation theorem (Cybenko 1989) that a two-layered design is suitable for the MLP. Because of this theorem, in this chapter, the two-layered network shown in Fig. 3.1 was chosen. The relationship between CPI,  $y$ , and economic variables,  $x$ , may be written as follows (Bishop 1995; Marwala 2012):

$$y_k = f_{outer} \left( \sum_{j=1}^M w_{kj}^{(2)} f_{inner} \left( \sum_{i=1}^d w_{ji}^{(1)} x_i + w_{j0}^{(1)} \right) + w_{k0}^{(2)} \right) \quad (3.1)$$

Here,  $w_{ji}^{(1)}$  and  $w_{ji}^{(2)}$  specify neural network weights in the first and second layers, respectively, going from input  $i$  to hidden unit  $j$ ,  $M$  is the number of hidden units,  $d$  is the number of output units while  $w_{j0}^{(1)}$  specifies the bias for the hidden unit  $j$ . This chapter uses a hyperbolic tangent function for the function  $f_{inner}(\bullet)$ . The function  $f_{outer}(\bullet)$  is linear because the problem we are handling is a regression problem (Bishop 1995; Marwala and Lagazio 2011).

On training regular neural networks, the network weights are identified while the training of probabilistic neural networks identifies the probability distributions of the network weights. An objective function must be selected to identify the weights in Eq. 3.1. An *objective function* is a mathematical representation of the global

objective of the problem (Marwala and Lagazio 2011). In this chapter, the principal objective was to identify a set of neural network weights, given the economic variables and the CPI and then rank the inputs in terms of importance.

If the training set  $D = \{x_k, y_k\}_{k=1}^N$  is used, where superscript  $N$  is the number of training examples, and assuming that the targets  $y$  are sampled independently given the  $k$ th inputs  $x_k$  and the weight parameters  $w_{kj}$  then the objective function,  $E$ , may be written using the sum of squares of errors objective function (Rosenblatt 1961; Bishop 1995; Wang and Lu 2006; Marwala 2009):

$$E_D = -\beta \sum_{n=1}^N \sum_{k=1}^K \{t_{nk} - y_{nk}\}^2 \quad (3.2)$$

Here,  $t_{nk}$  is the target vector for the  $n$ th output and  $k$ th training example,  $N$  is the number of training examples,  $K$  is the number of network output units,  $n$  is the index for the training pattern,  $\beta$  is the data contribution to the error, and  $k$  is the index for the output units.

The sum of squares error objective function was chosen because it has been found to be suited to regression problems than the cross-entropy error objective function (Bishop 1995). Equation 3.2 can be regularized by presenting extra information to the objective function which is penalty function to solve an ill-posed problem or to prevent over-fitting by safeguarding smoothness of the solution to reach a trade-off between complexity and accuracy using (Bishop 1995; Marwala and Lagazio 2011):

$$E_W = -\frac{\alpha}{2} \sum_{j=1}^W w_j^2 \quad (3.3)$$

Here,  $\alpha$  is the prior contribution to the regularization error and  $W$  is the number of network weights. This regularization parameter penalizes weights of large magnitudes (Bishop 1995; Tibshirani 1996; Marwala 2009; Marwala and Lagazio 2011). To solve for the weights in Eq. 3.1, the back-propagation technique described in the next section was used.

By merging Eqs. 3.2 and 3.3, the complete objective function can be written as follows (Bishop 1995):

$$\begin{aligned} E &= \beta E_D + \alpha E_W \\ &= -\beta \sum_{n=1}^N \sum_{k=1}^K \{t_{nk} - y_{nk}\}^2 - \frac{\alpha}{2} \sum_{j=1}^W w_j^2 \end{aligned} \quad (3.4)$$

### 3.2.1.1 Back-Propagation Method

As described by Marwala and Lagazio (2011), the back-propagation technique is a procedure of training neural networks (Bryson and Ho 1989; Rumelhart et al. 1986; Russell and Norvig 1995). Back-propagation is a supervised learning procedure and is basically an application of the Delta rule (Rumelhart et al. 1986; Russell and Norvig 1995). Back-propagation requires that the activation function (seen in Eq. 3.1) can be differentiated and is divided into the propagation and weight update. According to Bishop (1995) the propagation characteristic has these steps:

- the propagation's output activations is enacted by forward propagating the training pattern's input into the neural network,
- the deltas of all output and hidden neurons are estimated by back-propagating the output activations in the neural network by applying the training target.

The weight update characteristic applies the output delta and input activation to get the gradient of the weight. The weight in the reverse direction of the gradient is utilized by subtracting a proportion of it from the weight and then this proportion influences the performance of the learning process. The sign of the gradient of a weight defines the direction where the error increases and this is the motive that the weight is updated in the opposite direction (Bishop 1995), and this process is recurred until convergence.

In basic terms, back-propagation is used to identify the network weights given the training data, using an optimization technique. Generally, the weights can be identified using the following iterative technique (Werbos 1974; Zhao et al. 2010; Marwala and Lagazio 2011):

$$\{w\}_{i+1} = \{w\}_i - \eta \frac{\partial E}{\partial \{w\}} (\{w\}_i) \quad (3.5)$$

In Eq. 3.5, the parameter  $\eta$  is the learning rate while  $\{\}$  characterizes a vector. The minimization of the objective function,  $E$ , is attained by computing the derivative of the errors in Eq. 3.2 with respect to the network's weight. The derivative of the error is calculated with respect to the weight which joins the hidden layer to the output layer and may be written as follows, using the chain rule (Bishop 1995):

$$\begin{aligned} \frac{\partial E}{\partial w_{kj}} &= \frac{\partial E}{\partial a_k} \frac{\partial a_k}{\partial w_{kj}} \\ &= \frac{\partial E}{\partial y_k} \frac{\partial y_k}{\partial a_k} \frac{\partial a_k}{\partial w_{kj}} \\ &= \sum_n f'_{outer}(a_k) \frac{\partial E}{\partial y_{nk}} z_j \end{aligned} \quad (3.6)$$

In Eq. 3.6,  $z_j = f_{inner}(a_j)$  and  $a_k = \sum_{j=0}^M w_{kj}^{(2)} y_j$ . The derivative of the error with respect to weight which connects the hidden layer to the output layer may also be written using the chain rule (Bishop 1995):

$$\begin{aligned} \frac{\partial E}{\partial w_{kj}} &= \frac{\partial E}{\partial a_k} \frac{\partial a_k}{\partial w_{kj}} \\ &= \sum_n f'_{inner}(a_j) \sum_k w_{kj} f'_{outer}(a_k) \frac{\partial E}{\partial y_{nk}} \end{aligned} \quad (3.7)$$

In Eq. 3.7,  $a_j = \sum_{i=1}^d w_{ji}^{(1)} x_i$ . The derivative of the objective function in Eq. 3.2 may thus be written as (Bishop 1995):

$$\frac{\partial E}{\partial y_{nk}} = (t_{nk} - y_{nk}) \quad (3.8)$$

while that of the hyperbolic tangent function is (Bishop 1995):

$$f'_{inner}(a_j) = \sec h^2(a_j) \quad (3.9)$$

Now that it has been explained how to estimate the gradient of the error with respect to the network weights using back-propagation procedure, Eq. 3.6 can be applied to update the network weights by using an optimization process until some pre-defined stopping criterion is realized. If the learning rate in Eq. 3.5 is fixed, then this is called the *steepest descent optimization* technique (Robbins and Monro 1951). Conversely, the steepest descent technique is not computationally efficient and, hence, an improved scheme requires to be identified. In this chapter, the scaled conjugate gradient technique was applied (Møller 1993), the subject of the next section.

### 3.2.1.2 Scaled Conjugate Gradient Method

The method in which the network weights are approximated from the data is by using non-linear optimization technique (Mordecai 2003). In this chapter, the scaled conjugate gradient technique (Møller 1993) was used. As described earlier in this chapter, the weight vector, which provides the minimum error, is calculated by simulating sequential steps through the weight space as presented in Eq. 3.7 until a pre-determined stopping criterion is achieved. Different procedures select this learning rate in a different ways. In this section, the gradient descent technique is described, and after that how it is modified to the conjugate gradient technique (Hestenes and Stiefel 1952). For the gradient descent technique, the step size is defined as  $-\eta \partial E / \partial w$ , where the parameter  $\eta$  is the learning rate and the gradient of the error is estimated using the back-propagation method.



If the learning rate is adequately small, the value of the error decreases at each following step until a minimum value for the error between the model prediction and training target data is achieved. The drawback with this method is that it is computationally expensive when compared to other procedures. For the conjugate gradient method, the quadratic function of the error is minimized at every step over a gradually increasing linear vector space that comprises the global minimum of the error (Luenberger 1984; Fletcher 1987; Bertsekas 1995; Marwala 2012). In the conjugate gradient technique, the following steps are used (Haykin 1999; Marwala 2009; Babaie-Kafaki et al. 2010; Marwala and Lagazio 2011):

1. Select the initial weight vector  $\{w\}_0$ .
2. Estimate the gradient vector  $\frac{\partial E}{\partial \{w\}} (\{w\}_0)$ .
3. At each step,  $n$ , apply the line search to identify  $\eta(n)$  that minimizes  $E(\eta)$  indicating the objective function expressed in terms of  $\eta$  for fixed values of  $w$  and  $-\frac{\partial E}{\partial \{w\}} (\{w_n\})$ .
4. Evaluate that the Euclidean norm of the vector  $-\frac{\partial E}{\partial w} (\{w_n\})$  is sufficiently less than that of  $-\frac{\partial E}{\partial w} (\{w_0\})$ .
5. Change the weight vector using Eq. 3.4.
6. For  $w_{n+1}$ , calculate the changed gradient  $\frac{\partial E}{\partial \{w\}} (\{w\}_{n+1})$ .
7. Apply the Polak-Ribière technique to estimate:

$$\beta(n+1) = \frac{\nabla E(\{w\}_{n+1})^T (\nabla E(\{w\}_{n+1}) - \nabla E(\{w\}_n))}{\nabla E(\{w\}_n)^T \nabla E(\{w\}_n)}.$$

8. Change the direction vector

$$\frac{\partial E}{\partial \{w\}} (\{w\}_{n+2}) = \frac{\partial E}{\partial \{w\}} (\{w\}_{n+1}) - \beta(n+1) \frac{\partial E}{\partial \{w\}} (\{w\}_n).$$

9. Let  $n = n + 1$  and return to step 3.
10. Terminate when the following criterion is met:

$$\varepsilon = \frac{\partial E}{\partial \{w\}} (\{w\}_{n+2}) - \frac{\partial E}{\partial \{w\}} (\{w\}_{n+1}) \quad \text{where } \ll 0.$$

The scaled conjugate gradient technique is different from the conjugate gradient scheme because it does not include the line search referred to in Step 3. The step-size is estimated using the following formula (Møller 1993):

$$\begin{aligned} \eta(n) = 2 \left( \eta(n) - \left( \frac{\partial E(n)}{\partial \{w\}}(n) \right)^T H(n) \left( \frac{\partial E(n)}{\partial \{w\}}(n) \right) \right. \\ \left. + \eta(n) \left\| \left( \frac{\partial E(n)}{\partial \{w\}}(n) \right) \right\|^2 / \left\| \left( \frac{\partial E(n)}{\partial \{w\}}(n) \right) \right\|^2 \right) \end{aligned} \quad (3.10)$$

where  $H$  is the Hessian of the gradient. The scaled conjugate gradient technique is applied because it has been observed to resolve the optimization problems of training an MLP network more computationally effective than the gradient descent and conjugate gradient approaches (Bishop 1995).

### 3.2.2 Bayesian Framework

Multi-layered neural networks are parameterized classification models that make probabilistic assumptions about the data. The probabilistic outlook of these models is enabled by the application of the Bayesian framework (Marwala 2012). Learning algorithms are viewed as approaches for identifying parameter values that look probable in the light of the presented data. The learning process is performed by dividing the data into training, validation and testing sets. This is done for model selection and to ensure that the trained network is not biased towards the training data it has seen. Another way of realizing this is by the application of the regularization framework, which comes naturally from the Bayesian formulation and is now explained in detail in this chapter.

Thomas Bayes was the originator of the Bayes' theorem and Pierre-Simon Laplace generalized the theorem and it has been applied it to problems such as in engineering, statistics, political science and reliability (Stigler 1986; Fienberg 2006; Bernardo 2005; Marwala and Lagazio 2011). Initially, the Bayesian method applied uniform priors and was known as the "inverse probability" and later succeeded by a process called "frequentist statistics" also called the maximum-likelihood method. The maximum-likelihood method is intended to identify the most probable solution without concern to the probability distribution of that solution. The maximum-likelihood method technique is basically a special case of Bayesian results indicating the most probable solution in the distribution of the posterior probability function. Bayesian procedure consists of the following concepts (Bishop 1995):

- The practice of hierarchical models and the marginalization over the values of irrelevant parameters using methods such as the Markov chain Monte Carlo techniques.
- The iterative use of the Bayes' theorem as data points are acquired and after approximating a posterior distribution, the posterior equals the next prior.
- In the maximum-likelihood method, a hypothesis is a proposition which must be proven right or wrong while in a Bayesian procedure, a hypothesis has a probability.

The Bayesian method has been applied to many complex problems, including those of finite element model updating (Marwala and Sibisi 2005), missing data estimation (Marwala 2009), health risk assessment (Goulding et al. 2012), astronomy (Petremand et al. 2012), classification of file system activity (Khan 2012), simulating ecosystem metabolism (Shen and Sun 2012) and in image processing

(Thon et al. 2012). The problem of identifying the weights ( $w_i$ ) and biases (with subscripts 0 in Fig. 3.1) in the hidden layers may be posed in the Bayesian form as (Box and Tiao 1973; Marwala 2012):

$$P(\{w\} | [D]) = \frac{P([D] | \{w\}) P(\{w\})}{P([D])} \quad (3.11)$$

where  $P(w)$  is the probability distribution function of the weight-space in the absence of any data, also known as the prior distribution and  $D \equiv (y_1, \dots, y_N)$  is a matrix containing the data. The quantity  $P(w | D)$  is the posterior probability distribution after the data have been seen and  $P(D | w)$  is the likelihood function.

### 3.2.2.1 Likelihood Function

The likelihood function is the notion that expresses the probability of the model which depends on the weight parameters of a model to be true. It is fundamentally the probability of the observed data, given the free parameters of the model. The likelihood can be expressed mathematically as follows, by using the sum of squares error (Edwards 1972; Bishop 1995; Marwala 2012):

$$\begin{aligned} P([D] | \{w\}) &= \frac{1}{Z_D} \exp(-\beta E_D) \\ &= \frac{1}{Z_D} \exp\left(\beta \sum_n^N \sum_k^K \{t_{nk} - y_{nk}\}^2\right) \end{aligned} \quad (3.12)$$

In Eq. 3.12,  $E_D$  is the sum of squares of error function,  $\beta$  represents the hyper-parameters, and  $Z_D$  is a normalization constant which can be approximated as follows (Bishop 1995):

$$Z_D = \int_{-\infty}^{\infty} \exp\left(\beta \sum_n^N \sum_k^K \{t_{nk} - y_{nk}\}^2\right) d\{w\} \quad (3.13)$$

### 3.2.2.2 Prior Function

The *prior* probability distribution is the assumed probability of the free parameters and is approximated by a knowledgeable expert (Jaynes 1968; Bernardo 1979; Marwala and Lagazio 2011). There are many types of priors and these comprise informative and uninformative priors. An *informative prior* expresses accurate, particular information about a variable while an *uninformative prior* states general information about a variable. A prior distribution that assumes that model parameters are of the same order of magnitude can be written as follows (Bishop 1995):

$$\begin{aligned}
 P(\{w\}) &= \frac{1}{Z_w} \exp(-E_w) \\
 &= \frac{1}{Z_w} \exp\left(-\frac{\alpha}{2} \sum_j^W w_j^2\right)
 \end{aligned} \tag{3.14}$$

Parameter  $\alpha$  represents the hyper-parameters, and  $Z_w$  is the normalization constant which can be approximated as follows (Bishop 1995):

$$Z_w = \int_{-\infty}^{\infty} \exp\left(-\frac{\alpha}{2} \sum_j^W w_j^2\right) d\{w\} \tag{3.15}$$

The prior distribution of a Bayesian method is the regularization parameter in Eq. 3.2. Regularization includes presenting additional information to the objective function, through a penalty function to solve an ill-posed problem or to prevent over-fitting to guarantee the smoothness of the solution to balance complexity with accuracy.

### 3.2.2.3 Posterior Function

The posterior probability is the probability of the network weights given the observed data. It is a conditional probability assigned after the appropriate evidence is taken into account (Lee 2004). It is estimated by multiplying the likelihood function with the prior function and dividing it by a normalization function. By combining Eqs. 3.11 and 3.14, the posterior distribution can be expressed as follows (Bishop 1995):

$$P(w|D) = \frac{1}{Z_s} \exp\left(\beta \sum_n^N \sum_k^K \{t_{nk} - y_{nk}\}^2 - \frac{\alpha}{2} \sum_j^W w_j^2\right) \tag{3.16}$$

where

$$\begin{aligned}
 Z_E(\alpha, \beta) &= \int \exp(-\beta E_D - \alpha E_w) dw \\
 &= \left(\frac{2\pi}{\beta}\right)^{\frac{N}{2}} + \left(\frac{2\pi}{\alpha}\right)^{\frac{W}{2}}
 \end{aligned} \tag{3.17}$$

Training the network using a Bayesian method gives the probability distribution of the weights shown in Eq. 3.1. The Bayesian method penalizes highly complex models and can choose an optimal model (Bishop 1995).

### 3.2.3 Automatic Relevance Determination

As described by Marwala and Lagazio (2011), an automatic relevance determination method is built by associating the hyper-parameters of the prior with each input variable. This, consequently, necessitates Eq. 3.14 to be generalized to form (MacKay 1991, 1992):

$$E_W = \frac{1}{2} \sum_k \alpha_k \{w\}^T [I_k] \{w\} \quad (3.18)$$

Here, superscript  $T$  is the transpose,  $k$  is the weight group and  $[I]$  is the identity matrix. By using the generalized prior in Eq. 3.18, the posterior probability in Eq. 3.16 becomes (Bishop 1995):

$$\begin{aligned} P(\{w\} | [D], H_i) &= \frac{1}{Z_s} \exp \left( \beta \sum_n \{t_n - y(\{x\}_n)\}^2 - \frac{1}{2} \sum_k \alpha_k \{w\}^T [I_k] \{w\} \right) \\ &= \frac{1}{Z_E} \exp(-E(\{w\})) \end{aligned} \quad (3.19)$$

where

$$Z_E(\alpha, \beta) = \left( \frac{2\pi}{\beta} \right)^{\frac{N}{2}} + \prod_k \left( \frac{2\pi}{\alpha_k} \right)^{\frac{W_k}{2}} \quad (3.20)$$

Here,  $W_k$  is the number of weights in group  $k$ . The evidence can be written as follows (Bishop 1995):

$$\begin{aligned} p([D] | \alpha, \beta) &= \frac{1}{Z_D Z_W} \int \exp(-E(\{w\})) d\{w\} \\ &= \frac{Z_E}{Z_D Z_W} \\ &= \frac{\left( \frac{2\pi}{\beta} \right)^{\frac{N}{2}} + \prod_k \left( \frac{2\pi}{\alpha_k} \right)^{W_k/2}}{\left( \frac{2\pi}{\beta} \right)^{\frac{N}{2}} \prod_k \left( \frac{2\pi}{\alpha_k} \right)^{W_k/2}} \end{aligned} \quad (3.21)$$

The simultaneous estimation of the network weights and the hyper-parameters can be achieved using a number of ways including the use of Monte Carlo methods or any of its derivatives to maximize the posterior probability distribution. Another way of achieving this goal is to first maximize the log evidence and thus giving the following estimations for the hyper-parameters (Bishop 1995):

$$\beta^{MP} = \frac{N - \gamma}{2E_D(\{w\}^{MP})} \quad (3.22)$$

$$\alpha_k^{MP} = \frac{\gamma_k}{2E_{W_k}(\{w\}^{MP})} \quad (3.23)$$

where  $\gamma = \sum_k \gamma_k$ ,  $2E_{W_k} = \{w\}^T [I_k] \{w\}$  and

$$\gamma_k = \sum_j \left( \frac{\pi_j - \alpha_k}{\eta_j} ([V]^T [I_k] [V])_{jj} \right) \quad (3.24)$$

and  $\{w\}^{MP}$  is the weight vector at the maximum point and this is identified in this chapter using the scaled conjugate gradient method,  $\eta_j$  are the eigenvalues of  $[A]$ , and  $[V]$  are the eigenvalues such that  $[V]^T [V] = [I]$ . To estimate the relevance of each input variable, the  $\alpha_k^{MP}$ ,  $\beta^{MP}$ , and the following steps are followed (MacKay 1991):

1. Randomly choose the initial values for the hyper-parameters.
2. Train the neural network using the scaled conjugate gradient algorithm to minimize the objective function in Eq. 3.4 and thus identify  $\{w\}^{MP}$ .
3. Apply the evidence framework to estimate the hyper-parameters using Eqs. 3.22 and 3.23.
4. If not converged go to Step 2.

### 3.3 Applications of ARD in Inflation Modeling

In this chapter we apply the ARD to identify variables that drive inflation. Inflation is measured using a concept called Consumer Price Index (CPI). Artificial intelligence has been used in the past to model inflation. For example, Şahin et al. (2004) applied neural networks and cognitive mapping to model Turkey's inflation dynamics while Anderson et al. (2012) applied neural network to estimate the functional relationship between certain component sub-indexes and the CPI extracted decision rules from the network.

Binner et al. (2010) studied the influence of money on inflation forecasting. They applied recurrent neural networks and kernel recursive least squares regression to identify the best fitting U.S.A. inflation prediction models and compared these to a naïve random walk model. Their results demonstrated no correlation between monetary aggregates and inflation. McAdam and McNelis (2005) used neural networks and models based on Phillips-curve formulations to forecast inflation. This proposed models outperformed the best performing linear models. Cao et al. (2012)

applied linear autoregressive moving average model (ARMA) and neural networks for forecasting medical cost inflation rates. The results showed that the neural network model outperformed the ARMA.

Nakamura (2005) applied neural networks for predicting inflation and the results from a U.S.A. data demonstrated that neural networks outperformed univariate autoregressive model, while Binner et al. (2006) used a neural network and a Markov switching autoregressive (MS-AR) model to predict U.S.A. inflation and found that MS-AR model performed better than neural networks.

The CPI is a measure of inflation in an economy. It measures the changes in prices of a fixed pre-selected basket of goods. A basket of goods which is used for calculating the CPI in South Africa is as follows (Anonymous 2012):

1. Food and non-alcoholic beverages: bread and cereals, meat, fish, milk, cheese, eggs, oils, fats, fruit, vegetables, sugar, sweets, desserts, and other foods
2. Alcoholic beverages and tobacco
3. Clothing and footwear
4. Housing and utilities: rents, maintenance, water, electricity, and others
5. Household contents, equipment, and maintenance
6. Health: medical equipment, outpatient, and medical service
7. Transport
8. Communication
9. Recreation and culture
10. Education
11. Restaurants and hotels
12. Miscellaneous goods and services: personal care, insurance, and financial services.

This basket is weighed and the variation of prices of these goods is tracked from month to month and this is a basis for calculating inflation. It must be noted that there is normally a debate as to whether this basket of goods is appropriate. For example, in South Africa where there are two economies, one developed and formal and another informal and under-developed, there is always a debate on the validity of the CPI. This is even more important because the salary negotiations are based on the CPI.

In this chapter, we use the CPI data from 1992 to 2011 to model the relationship between economic variables and the CPI. These economic variables are listed in Table 3.1. They represent the performance of various aspect of the economy represented by 23 variables in the agriculture, manufacturing, mining, energy, construction, etc. A multi-layered perceptron neural network with 23 input variables, 12 hidden nodes, and 1 output representing the CPI is constructed. The ARD based MLP network is trained using the scaled conjugate gradient method and all these techniques were described earlier in the chapter. The results indicating the relevance of each variable is indicated in Table 3.1.

From Table 3.1, the following variables are deemed to be essential for modeling the CPI and these are mining, transport, storage and communication, financial

**Table 3.1** Automatic relevance with multi-layer perceptron and scaled conjugate gradient

Variable	Alpha	Inverse alpha	Relative weights
Agriculture fish forestry	14.4832	0.0690	1.79
Mining	5.7440	0.1741	4.51
Manufacturing	24.2071	0.0413	1.07
Electricity gas water	6.8551	0.1459	3.78
Construction	7.3717	0.1357	3.51
Retail and trade	15.0679	0.0664	1.72
Transport, storage and communication	2.3174	0.4315	11.18
Financial intermediation, insurance, real estate, and business services	0.9391	1.0648	27.59
Community, social, and personal services	0.4626	2.1616	56.00
Government services	7.2632	0.1377	3.57
Gross value added at basic prices	4.7935	0.2086	5.40
Taxes less subsidies on products	0.6467	1.5462	40.06
Affordability	1.0664	0.9377	24.29
Economic growth	4.0215	0.2487	6.44
Rand/USD exchange	25.8858	0.0386	1.00
Prime interest	5.5639	0.1797	4.66
Repo rate	5.5639	0.1797	4.66
Gross domestic product	0.2545	3.9287	101.78
Household consumption	0.4407	2.2692	58.79
Investment	0.5909	1.6924	43.84
Government consumption	7.5703	0.1321	3.42
Exports	20.8664	0.0479	1.24
Imports	5.9678	0.0386	1.00

intermediation, insurance, real estate and business services, community, social and personal services, gross value added at basic prices, taxes less subsidies on products, affordability, economic growth, repo rate, gross domestic product, household consumption, and investment.

It should be noted, however, that these results are purely based on the data set that was analyzed and the methodology that was used which is the ARD that is based on the MLP. These conclusions may change from one economy to another and from one methodology to another e.g. support vector machines instead of the MLP.

### 3.4 Conclusions

This chapter presented the Bayesian and the evidence frameworks to create an automatic relevance determination technique. The ARD method was then applied to determine the relevance of economic variables that are essential for driving the



consumer price index. It is concluded that for the data analyzed using the MLP based ARD technique, the variables driving the CPI are mining, transport, storage and communication, financial intermediation, insurance, real estate and business services, community, social and personal services, gross value added at basic prices, taxes less subsidies on products, affordability, economic growth, repo rate, gross domestic product, household consumption, and investment.

## References

- Alsharafat W (2013) Applying artificial neural network and extended classifier system for network intrusion detection. *Int Arab J Inf Technol* 10:art. no. 6-3011
- Anderson RG, Binner JM, Schmidt VA (2012) Connectionist-based rules describing the pass-through of individual goods prices into trend inflation in the United States. *Econ Lett* 117:174–177
- Anonymous (2012) CPI data <http://www.statssa.gov.za/>. Last accessed 03 Sept 2012
- Babaie-Kafaki S, Ghanbari R, Mahdavi-Amiri N (2010) Two new conjugate gradient methods based on modified secant equations. *J Comput Appl Math* 234:1374–1386
- Bernardo JM (1979) Reference posterior distributions for Bayesian inference. *J R Stat Soc* 41:113–147
- Bernardo JM (2005) Reference analysis. *Handb Stat* 25:17–90
- Bertsekas DP (1995) *Non-linear programming*. Athenas Scientific, Belmont
- Binner JM, Elger CT, Nilsson B, Tepper JA (2006) Predictable non-linearities in U.S. inflation. *Econ Lett* 93:323–328
- Binner JM, Tino P, Tepper J, Anderson R, Jones B, Kendall G (2010) Does money matter in inflation forecasting? *Physica A Stat Mech Its Appl* 389:4793–4808
- Bishop CM (1995) *Neural networks for pattern recognition*. Oxford University Press, Oxford
- Böck M, Ogishima S, Tanaka H, Kramer S, Kaderali L (2012) Hub-centered gene network reconstruction using automatic relevance determination. *PLoS ONE* 7:art. no. e35077
- Box GEP, Tiao GC (1973) *Bayesian inference in statistical analysis*. Wiley, Hoboken
- Browne A, Jakary A, Vinogradov S, Fu Y, Deicken RF (2008) Automatic relevance determination for identifying thalamic regions implicated in schizophrenia. *IEEE Trans Neural Netw* 19:1101–1107
- Bryson AE, Ho YC (1989) *Applied optimal control: optimization, estimation, and control*. Xerox College Publishing, Kentucky
- Cao Q, Ewing BT, Thompson MA (2012) Forecasting medical cost inflation rates: a model comparison approach. *Decis Support Syst* 53:154–160
- Cybenko G (1989) Approximations by superpositions of sigmoidal functions. *Math Control Signal Syst* 2:303–314
- Edwards AWF (1972) *Likelihood*. Cambridge University Press, Cambridge
- Fienberg SE (2006) When did Bayesian inference become “Bayesian”? *Bayesian Anal* 1:1–40
- Fletcher R (1987) *Practical methods of optimization*. Wiley, New York
- Fu Y, Browne A (2008) Investigating the influence of feature correlations on automatic relevance determination. In: *Proceedings of the international joint conference on neural Networks*, Hong Kong, 2008, pp 661–665
- Goulding R, Jayasuriya N, Horan E (2012) A Bayesian network model to assess the public health risk associated with wet weather sewer overflows discharging into waterways. *Water Res* 46:4933–4940
- Hassoun MH (1995) *Fundamentals of artificial neural networks*. MIT Press, Cambridge, MA
- Haykin S (1999) *Neural networks*. Prentice-Hall, Upper Saddle River

- Hestenes MR, Stiefel E (1952) Methods of conjugate gradients for solving linear systems. *J Res Nat Bur Stand* 6:409–436
- Huang Y, Beck JL, Wu S, Li H (2012) Stochastic optimization using automatic relevance determination prior model for Bayesian compressive sensing. In: Proceedings of SPIE – the international society for optical engineering, San Diego, 2012, art. no. 834837
- Hurwitz E, Marwala T (2011) Suitability of using technical indicators as potential strategies within intelligent trading systems. In: Proceedings of the IEEE international conference on systems, man, and cybernetics, Anchorage, 2011, pp 80–84
- Ikuta C, Uwate Y, Nishio Y (2012) Multi-layer perceptron with positive and negative pulse glial chain for solving two-spirals problem. In: Proceedings of the international joint conference on neural networks, Brisbane, 2012, art. no. 6252725
- Jacobs JP (2012) Bayesian support vector regression with automatic relevance determination kernel for modeling of antenna input characteristics. *IEEE Trans Antenna Propag* 60:2114–2118
- Jaynes ET (1968) Prior probabilities. *IEEE Trans Syst Sci Cybern* 4:227–241
- Khan MNA (2012) Performance analysis of Bayesian networks and neural networks in classification of file system activities. *Comput Secur* 31:391–401
- Khoza M, Marwala T (2012) Computational intelligence techniques for modelling an economic system. In: Proceedings of the international joint conference on neural networks, Brisbane, 2012, pp 1–5
- Lee PM (2004) Bayesian statistics, an introduction. Wiley, Hoboken
- Leke B, Marwala T (2005) Optimization of the stock market input time-window using Bayesian neural networks. In: Proceedings of the IEEE international conference on service operations, logistics and informatics, Beijing, 2005, pp 883–894
- Leke B, Marwala T, Tettey T (2007) Using inverse neural network for HIV adaptive control. *Int J Comput Intell Res* 3:11–15
- Li X, Li L (2012) IP core based hardware implementation of multi-layer perceptrons on FPGAs: a parallel approach. *Adv Mater Res* 433–440:5647–5653
- Lisboa PJG, Etchells TA, Jarman IH, Arsene CTC, Aung MSH, Eleuteri A, Taktak AFG, Ambrogio F, Boracchi P, Biganzoli E (2009) Partial logistic artificial neural network for competing risks regularized with automatic relevance determination. *IEEE Trans Neural Netw* 20:1403–1416
- Luenberger DG (1984) Linear and non-linear programming. Addison-Wesley, Reading
- Lunga D, Marwala T (2006) Online forecasting of stock market movement direction using the improved incremental algorithm. *Lect Note Comput Sci* 4234:440–449
- MacKay DJC (1991) Bayesian methods for adaptive models. Ph.D. thesis, California Institute of Technology, Pasadena
- MacKay DJC (1992) A practical Bayesian framework for back propagation networks. *Neural Comput* 4:448–472
- Martínez-Rego D, Fontenla-Romero O, Alonso-Betanzos A (2012) Nonlinear single layer neural network training algorithm for incremental, nonstationary and distributed learning scenarios. *Pattern Recognit* 45:4536–4546
- Marwala T (2000) On damage identification using a committee of neural networks. *J Eng Mech* 126:43–50
- Marwala T (2001) Probabilistic fault identification using a committee of neural networks and vibration data. *J Aircr* 38:138–146
- Marwala T (2003) Fault classification using pseudo modal energies and neural networks. *Am Inst Aeronaut Astronaut J* 41:82–89
- Marwala T (2009) Computational intelligence for missing data imputation, estimation and management: knowledge optimization techniques. IGI Global Publications, New York
- Marwala T (2012) Condition monitoring using computational intelligence methods. Springer, London
- Marwala T, Hunt HEM (1999) Fault identification using finite element models and neural networks. *Mech Syst Signal Process* 13:475–490
- Marwala T, Lagazio M (2011) Militarized conflict modeling using computational intelligence techniques. Springer, London

- Marwala T, Sibisi S (2005) Finite element model updating using Bayesian framework and modal properties. *J Aircr* 42:275–278
- McAdam P, McNelis P (2005) Forecasting inflation with thick models and neural networks. *Econ Model* 22:848–867
- Meena K, Subramaniam K, Gomathy M (2013) Gender classification in speech recognition using fuzzy logic and neural network. *Int Arab J Inf Technol* 10:art. no. 4476-7
- Mohamed N, Rubin D, Marwala T (2006) Detection of epileptiform activity in human EEG signals using Bayesian neural networks. *Neural Inf Process Lett Rev* 10:1–10
- Møller AF (1993) A scaled conjugate gradient algorithm for fast supervised learning. *Neural Netw* 6:525–533
- Mordecia A (2003) Non-linear programming: analysis and methods. Dover Publishing, New York
- Mørupa M, Hansena LK (2009) Automatic relevance determination for multi-way models. *J Chemom* 23:352–363
- Nakamura E (2005) Inflation forecasting using a neural network. *Econ Lett* 86:1–8
- Nasir AA, Mashor MY, Hassan R (2013) Classification of acute leukaemia cells using multilayer perceptron and simplified fuzzy ARTMAP neural networks. *Int Arab J inf Technol* 10:art. no. 4626-12
- Nummenmaa A, Auranen T, Hämäläinen MS, Jääskeläinen IP, Sams M, Vehtari A, Lampinen J (2007) Automatic relevance determination based hierarchical Bayesian MEG inversion in practice. *Neuroimage* 37:876–889
- Oh CK, Beck JL, Yamada M (2008) Bayesian learning using automatic relevance determination prior with an application to earthquake early warning. *J Eng Mech* 134:1013–1020
- Patel P, Marwala T (2006) Neural networks, fuzzy inference systems and adaptive-neuro fuzzy inference systems for financial decision making. *Lect Note Comput Sci* 4234:430–439
- Petremand M, Jalobeanu A, Collet C (2012) Optimal bayesian fusion of large hyperspectral astronomical observations. *Stat Methodol* 9:1572–3127
- Prakash G, Kulkarni M, Sripathi Acharya U, Kalyanpur MN (2012) Classification of FSO channel models using radial basis function neural networks and their performance with luby transform codes. *Int J Artif Intell* 9:67–75
- Rezaeian-Zadeh M, Zand-Parsa S, Abghari H, Zolghadr M, Singh VP (2012) Hourly air temperature driven using multi-layer perceptron and radial basis function networks in arid and semi-arid regions. *Theor Appl Climatol* 109:519–528
- Robbins H, Monro S (1951) A stochastic approximation method. *Ann Math Stat* 22:400–407
- Rosenblatt F (1961) Principles of neurodynamics: perceptrons and the theory of brain mechanisms. Spartan, Washington DC
- Rumelhart DE, Hinton GE, Williams RJ (1986) Parallel distributed processing: explorations in the microstructure of cognition. MIT Press, Cambridge, MA
- Russell S, Norvig P (1995) Artificial intelligence: a modern approach. Prentice Hall, Englewood Cliffs
- Russell MJ, Rubin DM, Wigdorowitz B, Marwala T (2008) The artificial larynx: a review of current technology and a proposal for future development. *Proc Int Fed Med Biol Eng* 20:160–163
- Russell MJ, Rubin DM, Wigdorowitz B, Marwala T (2009a) Pattern recognition and feature selection for the development of a new artificial larynx. In: Proceedings of the 11th world congress on medical physics and biomedical engineering, Munich, 2009, pp 736–739
- Russell MJ, Rubin DM, Marwala T, Wigdorowitz B (2009b) A voting and predictive neural network system for use in a new artificial larynx. *Proc IEEE ICBPE*. doi:10.1109/ICBPE.2009.5384105
- Şahin ŞÖ, Ülengin FN, Ülengin B (2004) Using neural networks and cognitive mapping in scenario analysis: the case of Turkey's inflation dynamics. *Eur J Oper Res* 158:124–145
- Sanz J, Perera R, Huerta C (2012) Gear dynamics monitoring using discrete wavelet transformation and multi-layer perceptron neural networks. *Appl Soft Comput J* 12:2867–2878
- Shaltaf S, Mohammad A (2013) A hybrid neural network and maximum likelihood based estimation of chirp signal parameters. *Int Arab J Inf Technol* 10:art. no. 4580-12

- Shen X, Sun T (2012) Applications of bayesian modeling to simulate ecosystem metabolism in response to hydrologic alteration and climate change in the Yellow River Estuary, China. *Procedia Environ Sci* 13:790–796
- Shutin D, Buchgraber T, Kulkarni SR, Poor HV (2011) Fast variational sparse Bayesian learning with automatic relevance determination for superimposed signals. *IEEE Trans Signal Process* 59:6257–6261
- Shutin D, Kulkarni SR, Poor HV (2012) Incremental reformulated automatic relevance determination. *IEEE Trans Signal Process* 60:4977–4981
- Sinha K, Chowdhury S, Saha PD, Datta S (2013) Modeling of microwave-assisted extraction of natural dye from seeds of *Bixa orellana* (Annatto) using response surface methodology (RSM) and artificial neural network (ANN). *Ind Crop Prod* 41:165–171
- Smyrnakis MG, Evans DJ (2007) Classifying ischemic events using a Bayesian inference multilayer perceptron and input variable evaluation using automatic relevance determination. *Comput Cardiol* 34:305–308
- Stigler SM (1986) *The history of statistics*. Harvard University Press, Cambridge, MA
- Taspinar N, Isik Y (2013) Multiuser detection with neural network MAI detector in CDMA systems for AWGN and Rayleigh fading asynchronous channels. *Int Arab J Inf Technol* 10:art. no. 4525-5
- Thon K, Rue H, Skrøvseth SO, Godtliebsen F (2012) Bayesian multiscale analysis of images modeled as Gaussian Markov random fields. *Comput Stat Data Anal* 56:49–61
- Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J R Stat Soc* 58:267–288
- Ulusoy I, Bishop CM (2006) Automatic relevance determination for the estimation of relevant features for object recognition. In: *Proceedings of the IEEE 14th signal processing and communication applications, Antalya, 2006*, pp 1–4
- Valdés JJ, Romero E, Barton AJ (2012) Data and knowledge visualization with virtual reality spaces, neural networks and rough sets: application to cancer and geophysical prospecting data. *Expert Syst Appl* 39:13193–13201
- Van Calster B, Timmerman D, Nabney IT, Valentin L, Van Holsbeke C, Van Huffel S (2006) Classifying ovarian tumors using Bayesian multi-layer perceptrons and automatic relevance determination: a multi-center study. *Proc Eng Med Biol Soc* 1:5342–5345
- Vilakazi BC, Marwala T (2007) Condition monitoring using computational intelligence. In: Laha D, Mandal P (eds) *Handbook on computational intelligence in manufacturing and production management*, illustrated edn. IGI Publishers, New York
- Wang D, Lu WZ (2006) Interval estimation of urban ozone level and selection of influential factors by employing automatic relevance determination model. *Chemosphere* 62:1600–1611
- Werbos PJ (1974) *Beyond regression: new tool for prediction and analysis in the behavioral sciences*. Ph.D. thesis, Harvard University, Cambridge
- Wu D (2012) An improved multi-layer perceptron neural network for scattered point data surface reconstruction. *ICIC Express Lett Part B Appl* 3:41–46
- Wu W, Chen Z, Gao S, Brown EN (2010) Hierarchical Bayesian modeling of inter-trial variability and variational Bayesian learning of common spatial patterns from multichannel EEG. In: *Proceedings of the 2010 IEEE international conference on acoustics speech and signal processing, Dallas, 2010*, pp 501–504
- Yoon Y, Peterson LL (1990) Artificial neural networks: an emerging new technique. In: *Proceedings of the ACM SIGBDP conference on trends and directions in expert systems*, Cambridge, 1990, pp 417–422
- Zhang X, Gou L, Hou B, Jiao L (2010) Gaussian process classification using automatic relevance determination for SAR target recognition. In: *Proceedings of SPIE – the international society for optical engineering*, art. no. 78300R
- Zhao Z, Xin H, Ren Y, Guo X (2010) Application and comparison of BP neural network algorithm in MATLAB. In: *Proceedings of the international conference on measurement technology and mechatron automat*, New York, 2010, pp 590–593