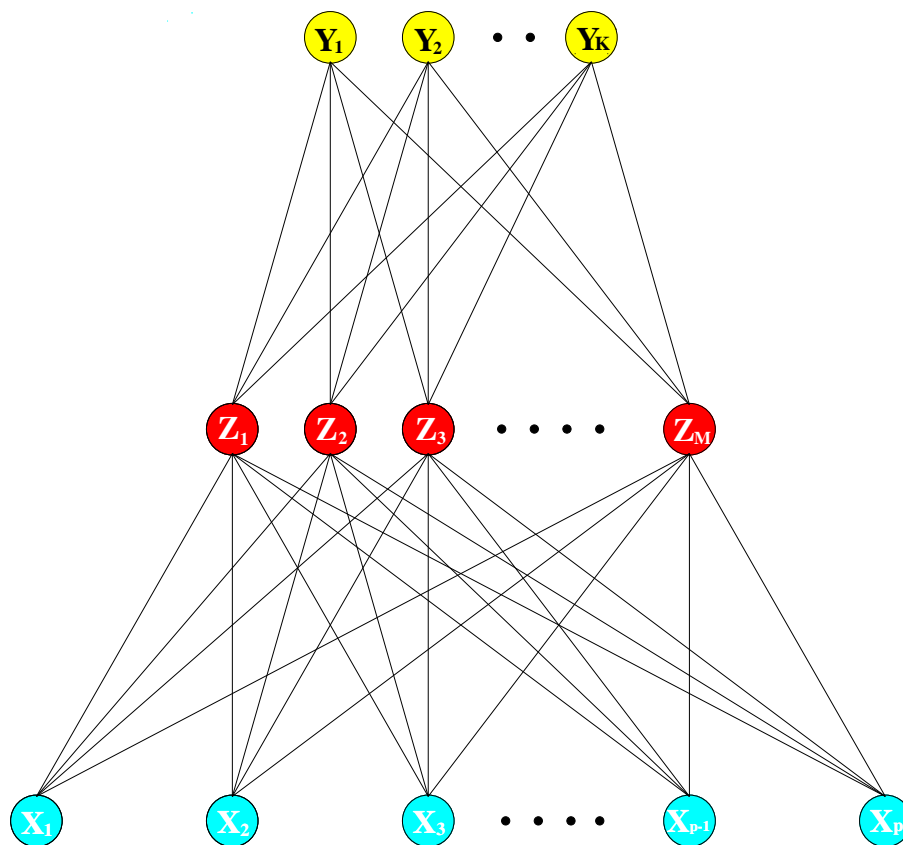


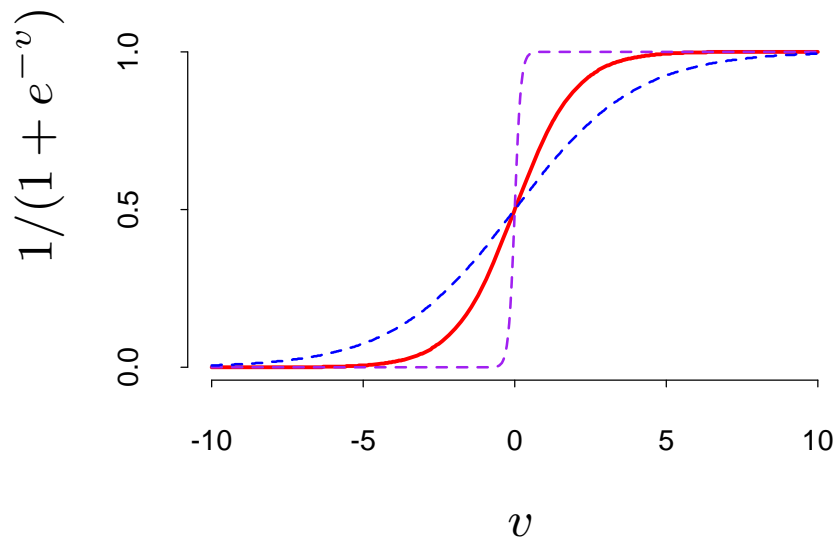
**FIGURE 11.1.** *Perspective plots of two ridge functions.*

(Left:)  $g(V) = 1/[1 + \exp(-5(V - 0.5))]$ , where  $V = (X_1 + X_2)/\sqrt{2}$ .

(Right:)  $g(V) = (V + 0.1) \sin(1/(V/3 + 0.1))$ , where  $V = X_1$ .

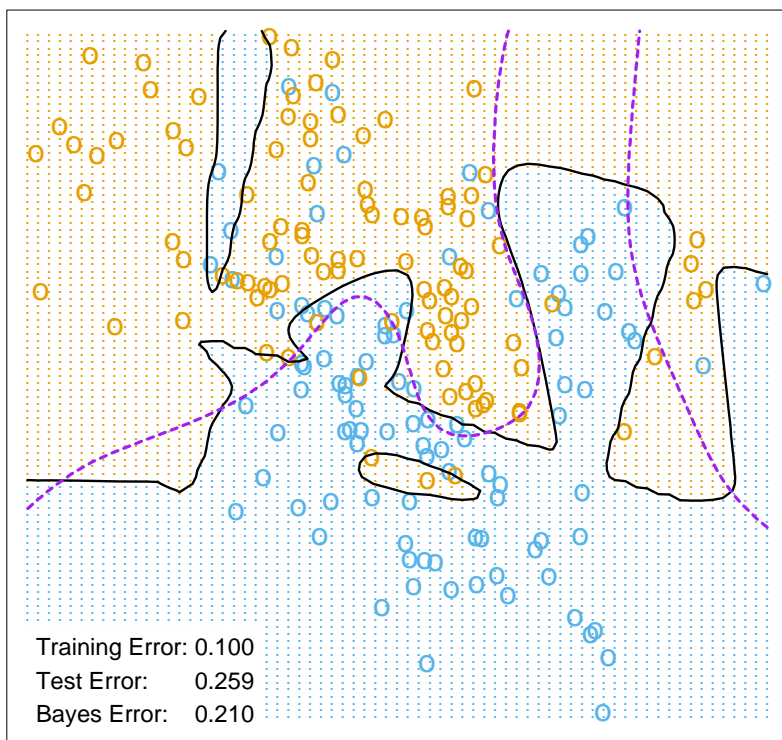


**FIGURE 11.2.** *Schematic of a single hidden layer, feed-forward neural network.*

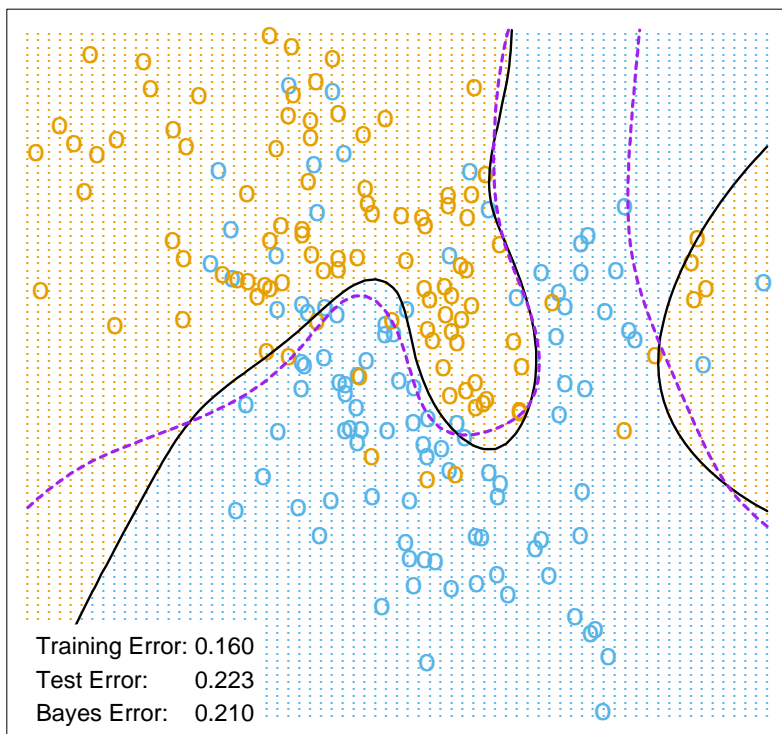


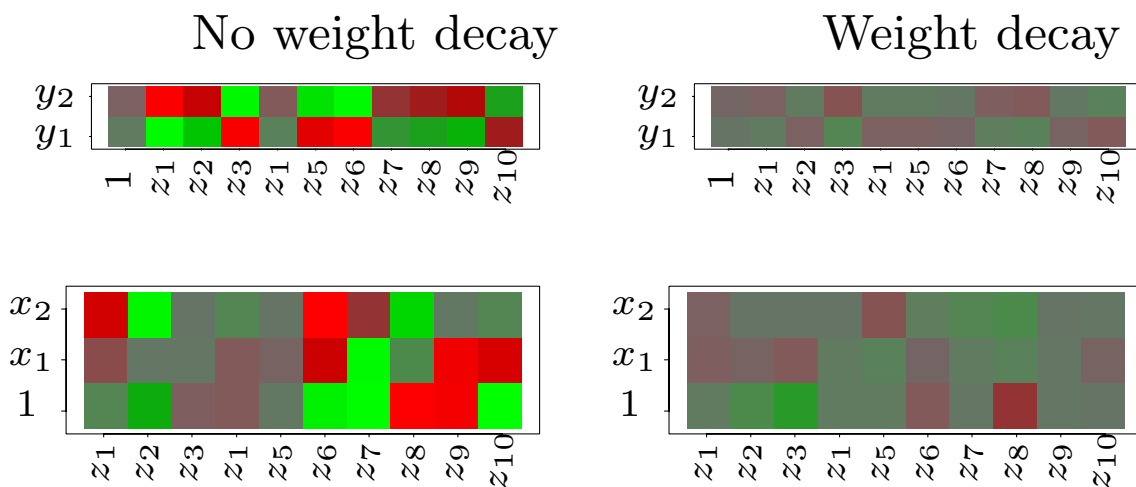
**FIGURE 11.3.** *Plot of the sigmoid function  $\sigma(v) = 1/(1 + \exp(-v))$  (red curve), commonly used in the hidden layer of a neural network. Included are  $\sigma(sv)$  for  $s = \frac{1}{2}$  (blue curve) and  $s = 10$  (purple curve). The scale parameter  $s$  controls the activation rate, and we can see that large  $s$  amounts to a hard activation at  $v = 0$ . Note that  $\sigma(s(v - v_0))$  shifts the activation threshold from 0 to  $v_0$ .*

Neural Network - 10 Units, No Weight Decay

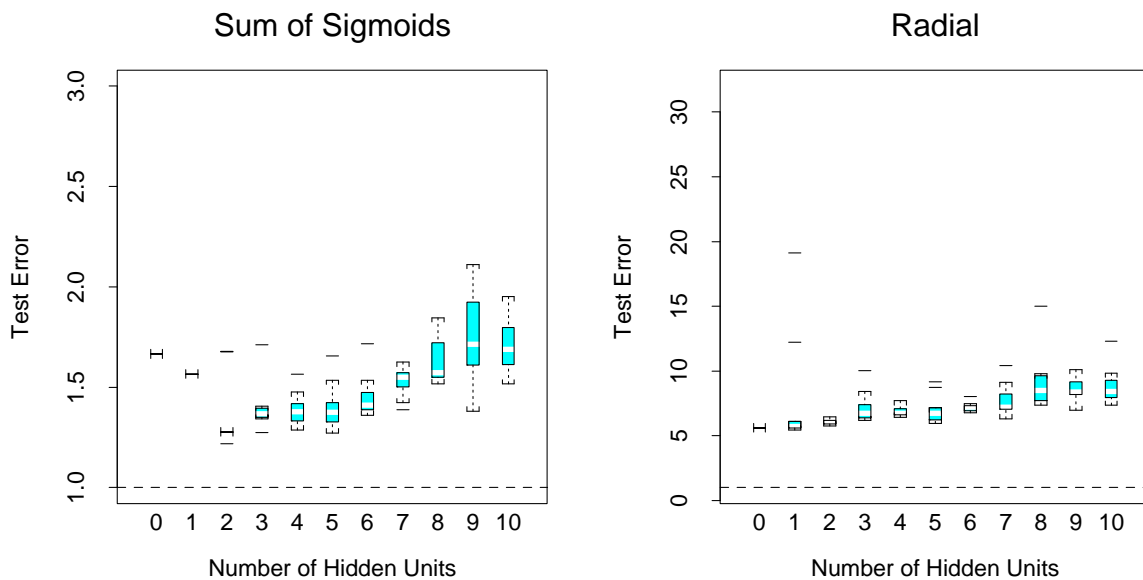


Neural Network - 10 Units, Weight Decay=0.02

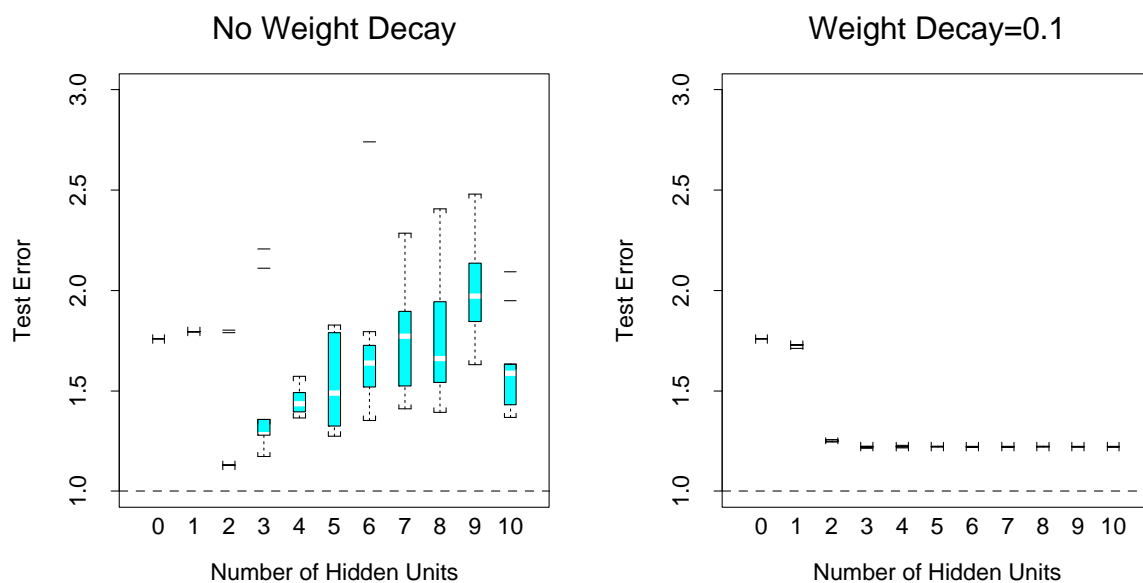




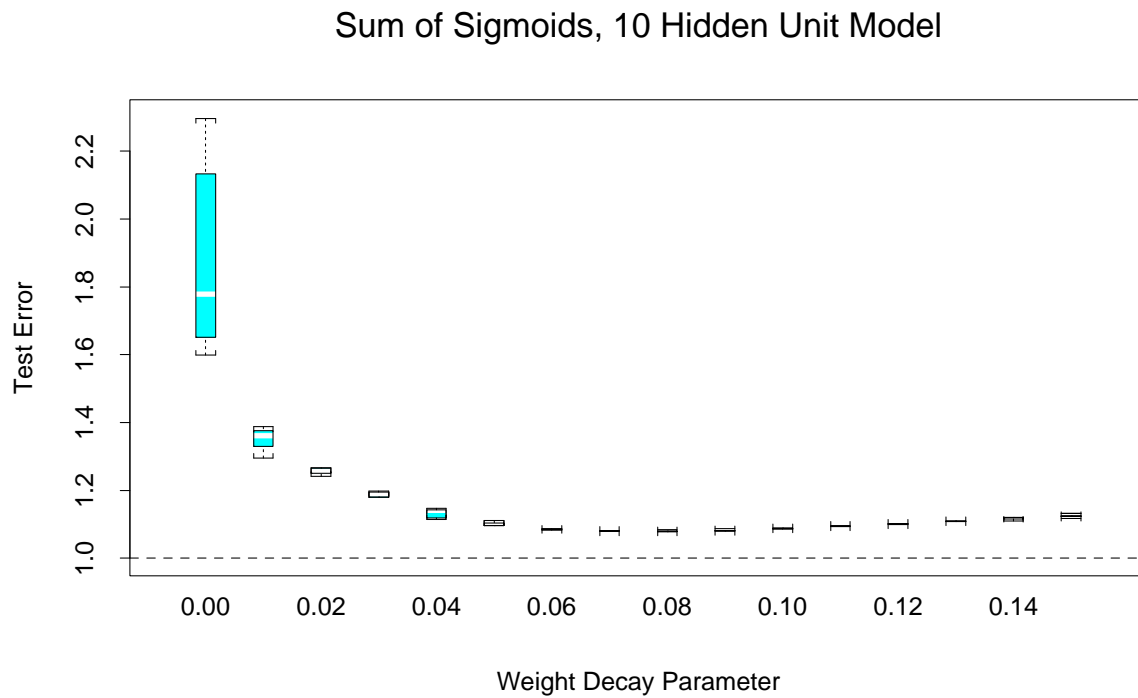
**FIGURE 11.5.** Heat maps of the estimated weights from the training of neural networks from Figure 11.4. The display ranges from bright green (negative) to bright red (positive).



**FIGURE 11.6.** *Boxplots of test error, for simulated data example, relative to the Bayes error (broken horizontal line). True function is a sum of two sigmoids on the left, and a radial function is on the right. The test error is displayed for 10 different starting weights, for a single hidden layer neural network with the number of units as indicated.*

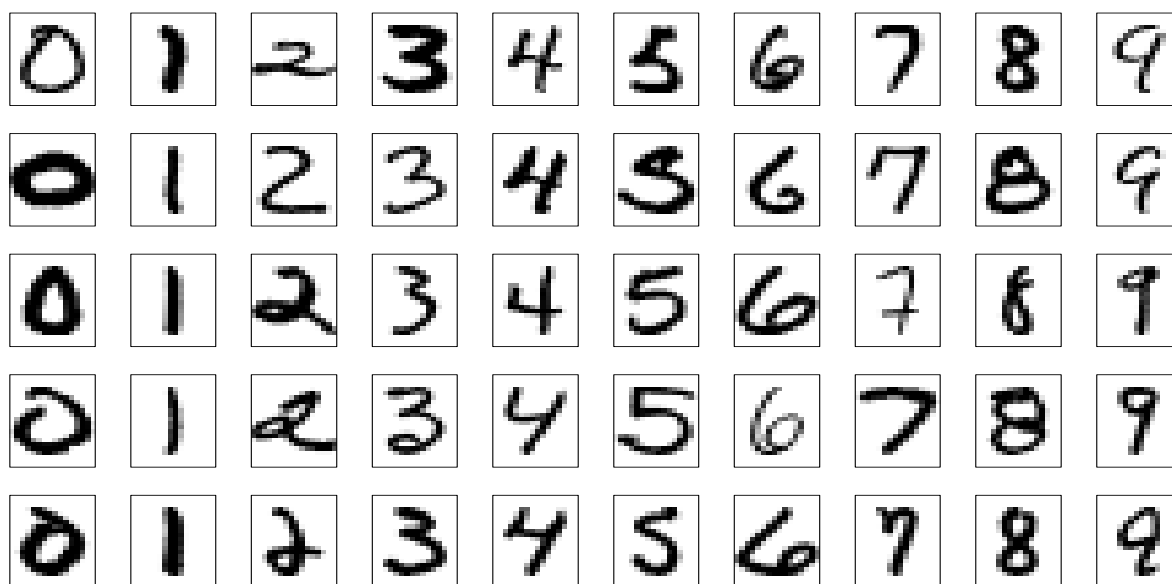


**FIGURE 11.7.** *Boxplots of test error, for simulated data example, relative to the Bayes error. True function is a sum of two sigmoids. The test error is displayed for ten different starting weights, for a single hidden layer neural network with the number units as indicated. The two panels represent no weight decay (left) and strong weight decay  $\lambda = 0.1$  (right).*

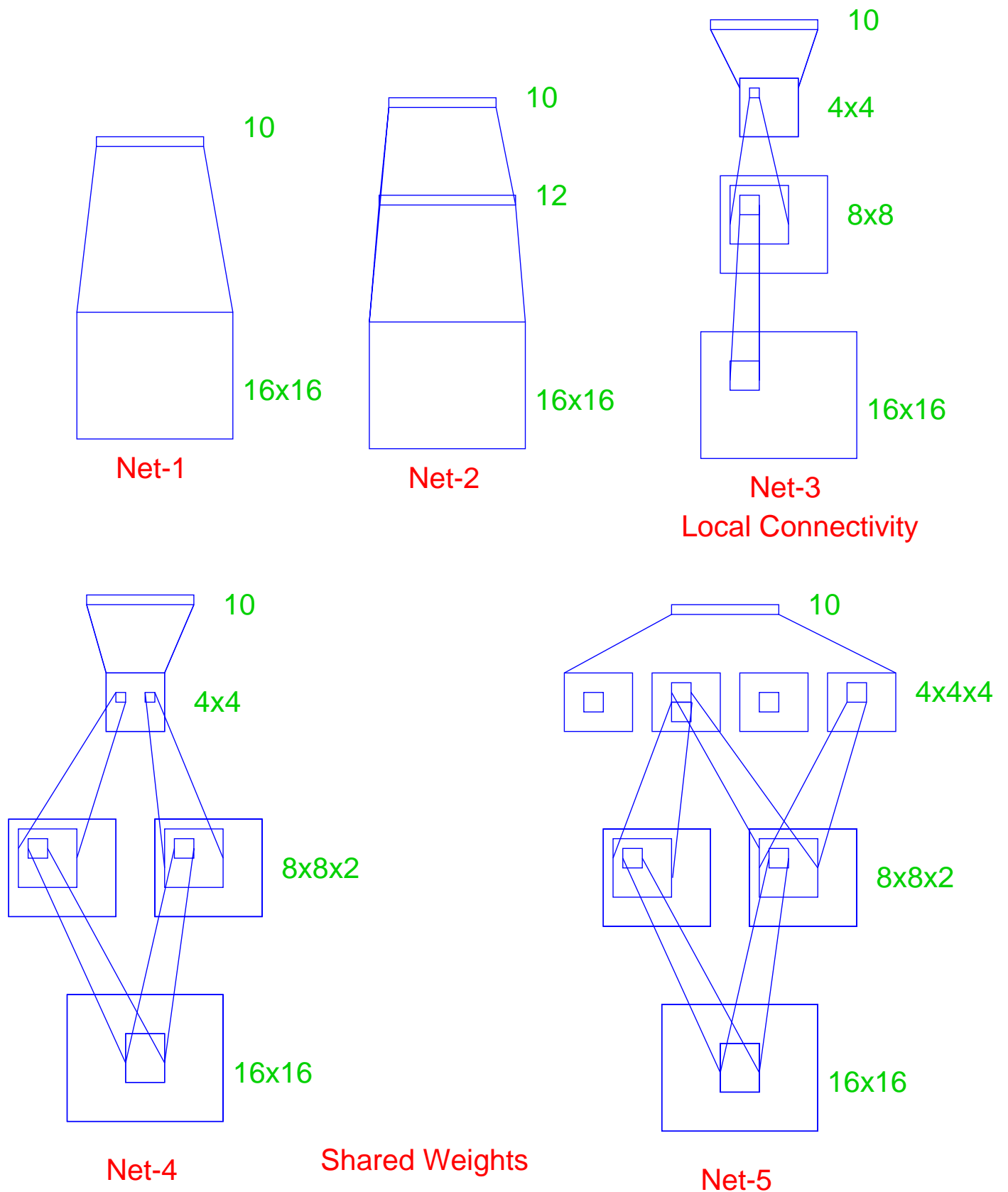


**FIGURE 11.8.** *Boxplots of test error, for simulated data example. True function is a sum of two sigmoids. The test error is displayed for ten different starting weights, for a single hidden layer neural network with ten hidden units and weight decay parameter value as indicated.*

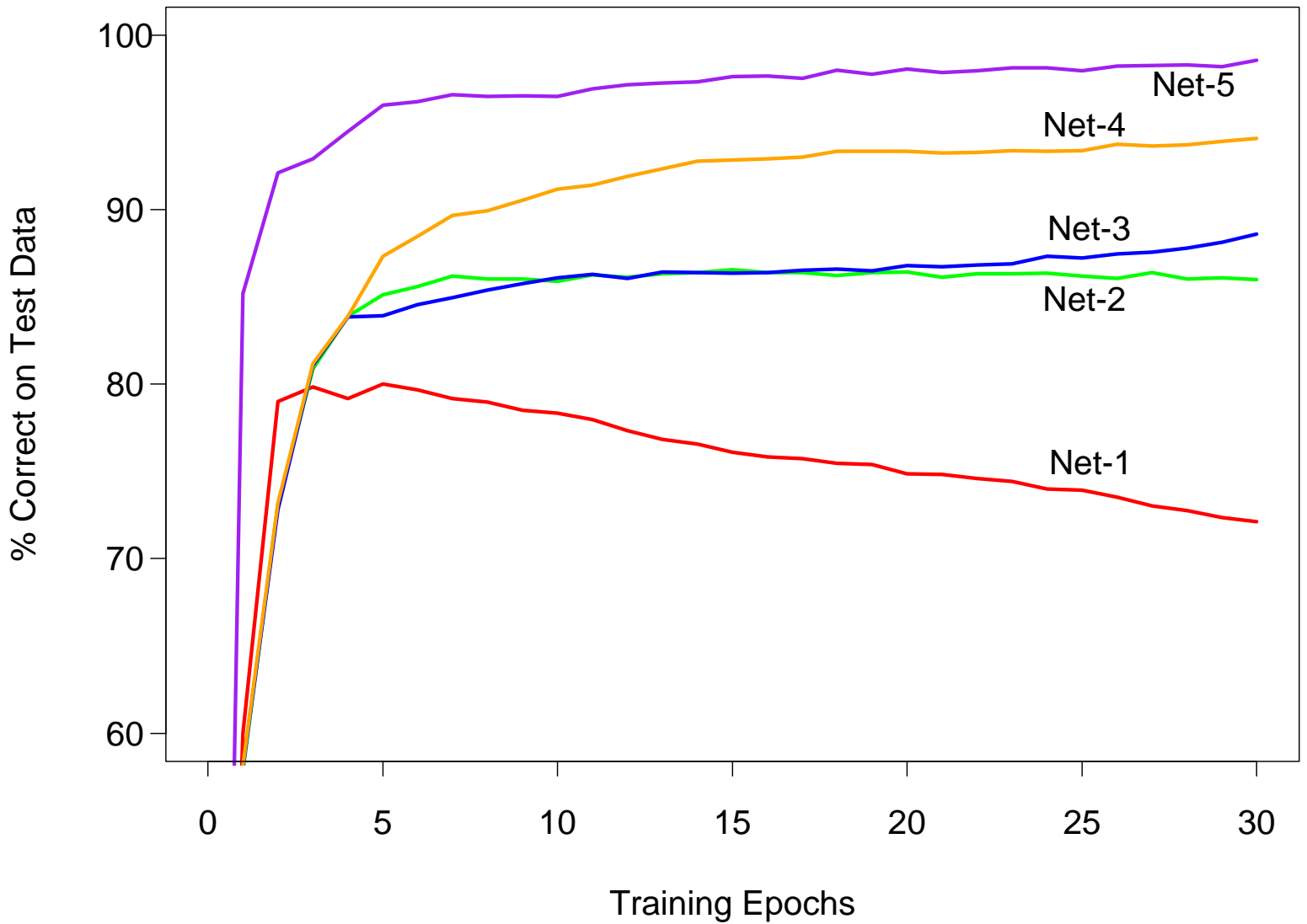




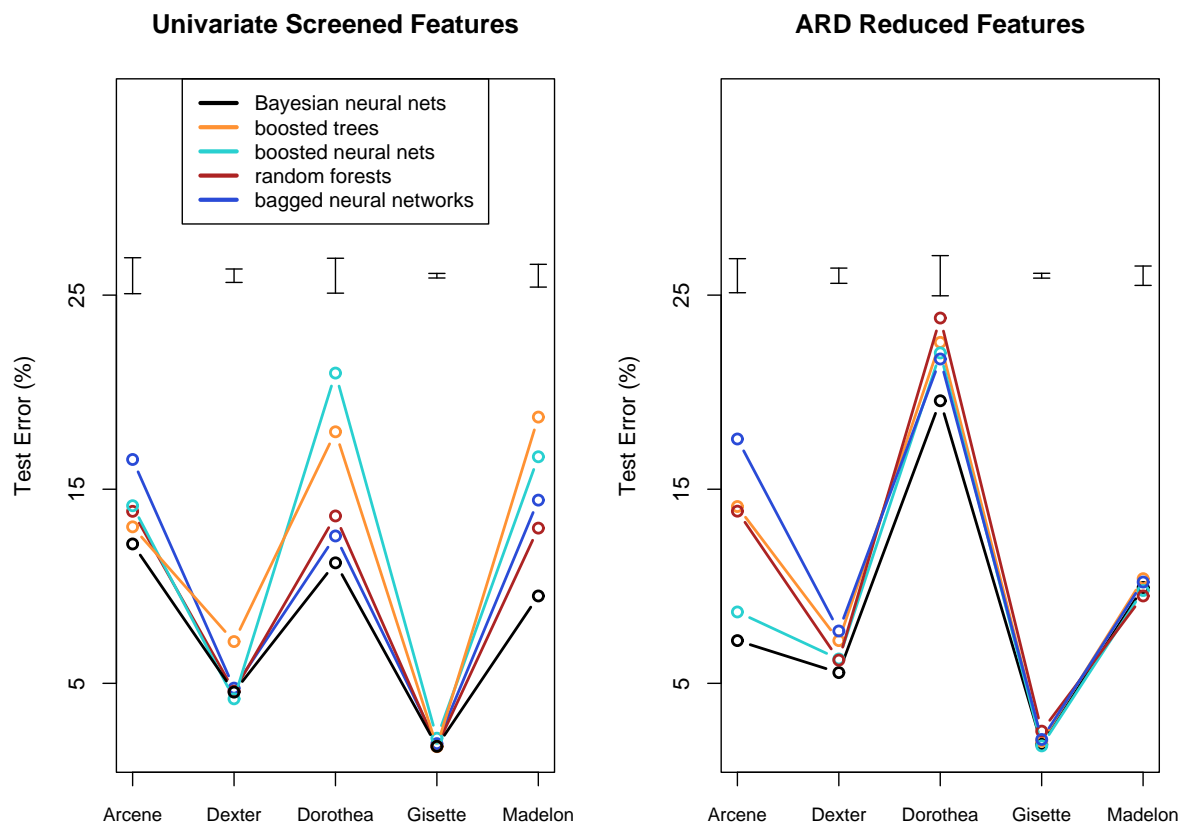
**FIGURE 11.9.** *Examples of training cases from ZIP code data. Each image is a  $16 \times 16$  8-bit grayscale representation of a handwritten digit.*



**FIGURE 11.10.** Architecture of the five networks used in the ZIP code example.



**FIGURE 11.11.** *Test performance curves, as a function of the number of training epochs, for the five networks of Table 11.1 applied to the ZIP code data. (?)*



**FIGURE 11.12.** Performance of different learning methods on five problems, using both univariate screening of features (top panel) and a reduced feature set from automatic relevance determination. The error bars at the top of each plot have width equal to one standard error of the difference between two error rates. On most of the problems several competitors are within this error bound.