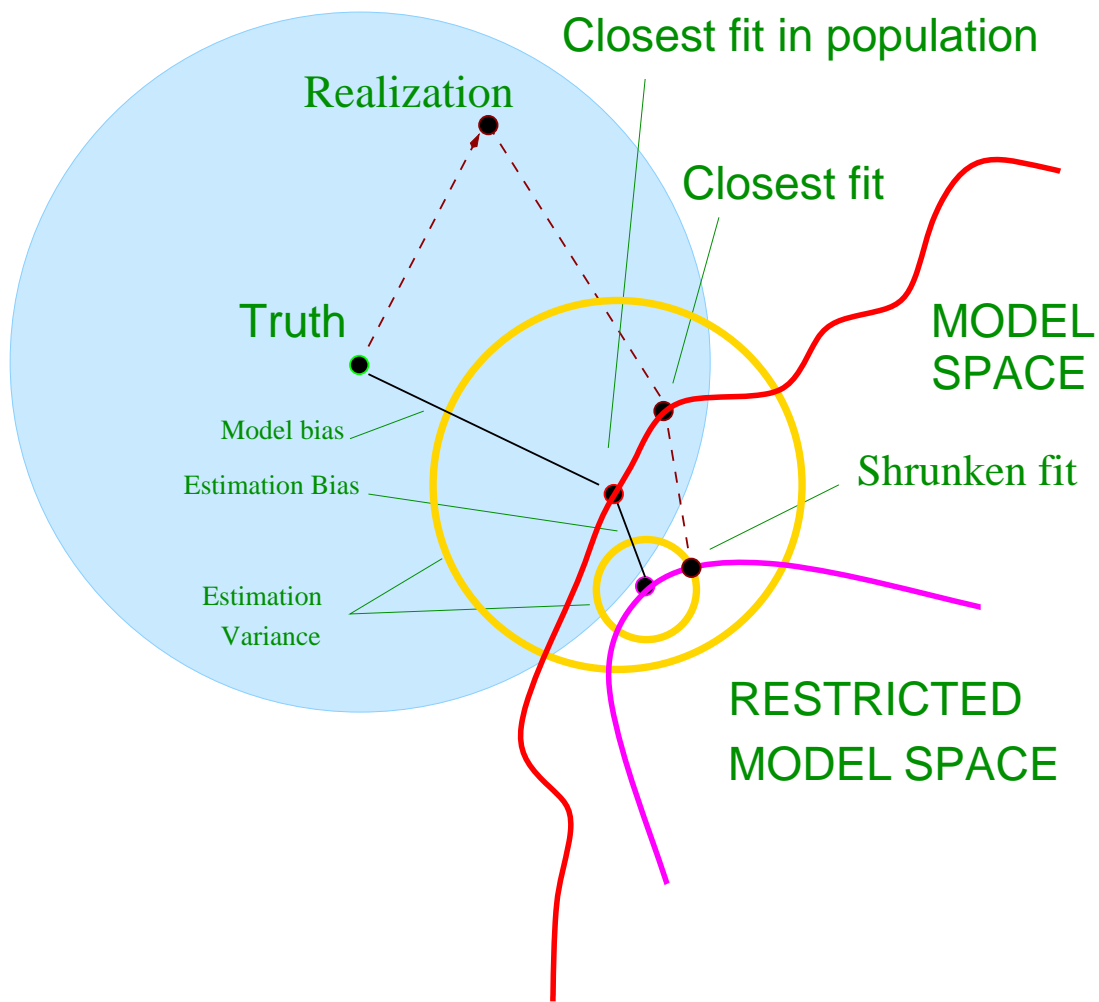
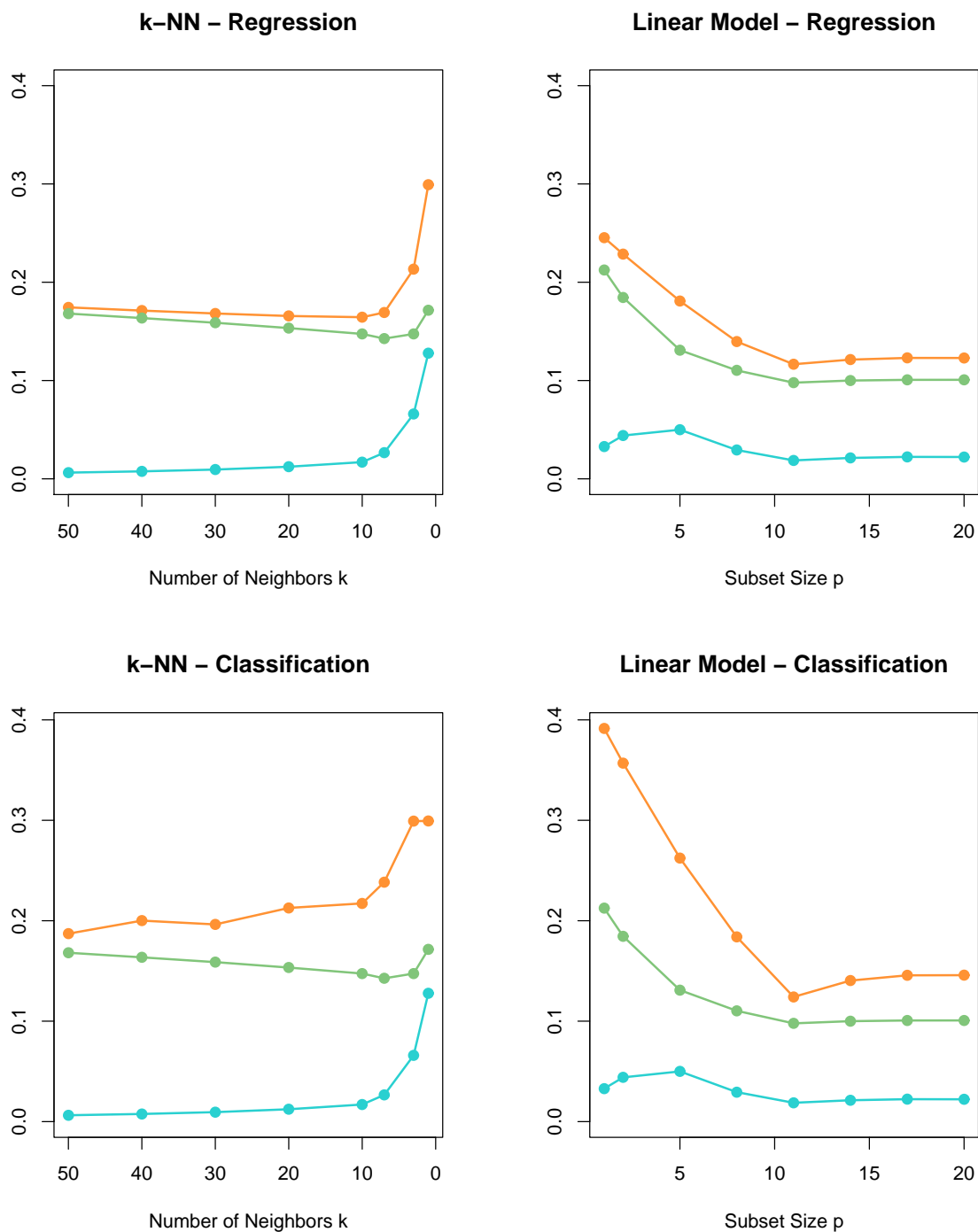


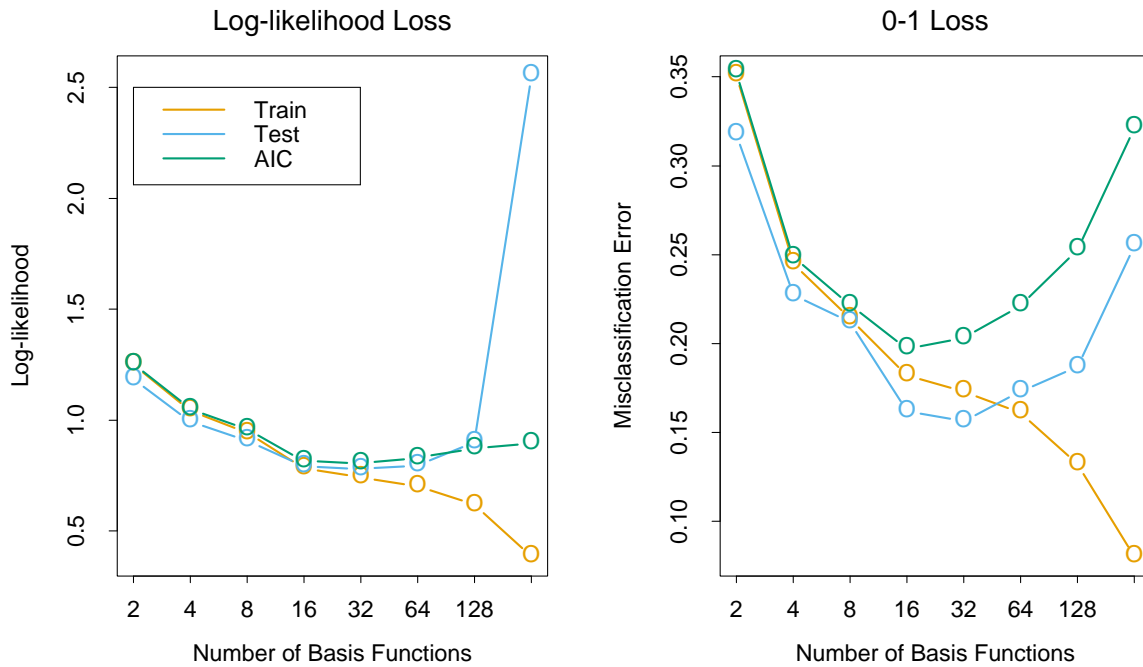
**FIGURE 7.1.** Behavior of test sample and training sample error as the model complexity is varied. The light blue curves show the training error  $\overline{\text{err}}$ , while the light red curves show the conditional test error  $\text{Err}_{\mathcal{T}}$  for 100 training sets of size 50 each, as the model complexity is increased. The solid curves show the expected test error  $\text{Err}$  and the expected training error  $\text{E}[\overline{\text{err}}]$ .



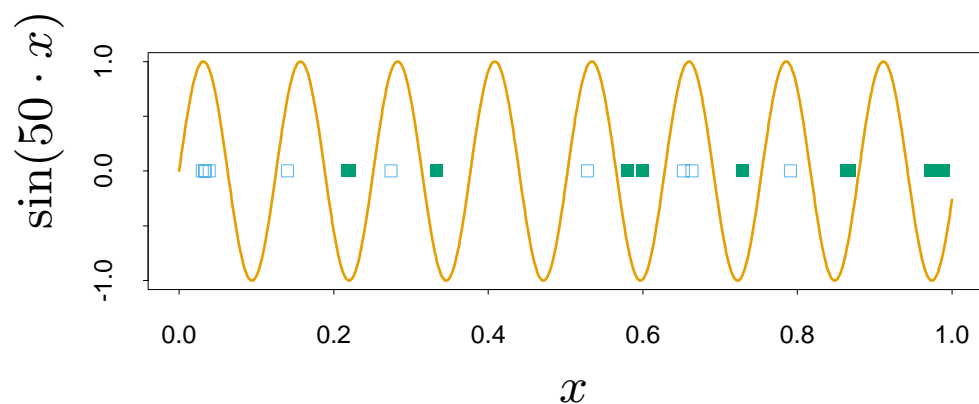
**FIGURE 7.2.** Schematic of the behavior of bias and variance. The model space is the set of all possible predictions from the model, with the “closest fit” labeled with a black dot. The model bias from the truth is shown, along with the variance, indicated by the large yellow circle centered at the black dot labeled “closest fit in population.” A shrunken or regularized fit is also shown, having additional estimation bias, but smaller prediction error due to its decreased variance.



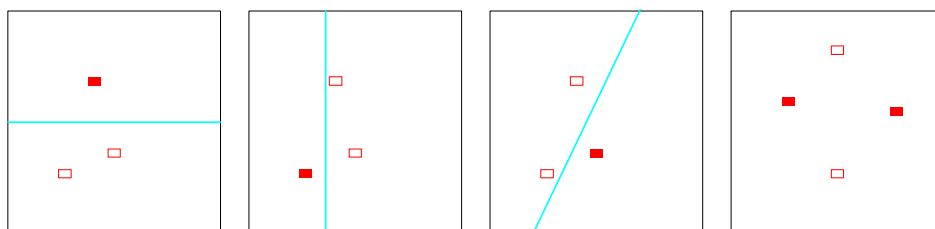
**FIGURE 7.3.** *Expected prediction error (orange), squared bias (green) and variance (blue) for a simulated example. The top row is regression with squared error loss; the bottom row is classification with 0–1 loss. The models are  $k$ -nearest neighbors (left) and best subset regression of size  $p$  (right). The variance and bias*



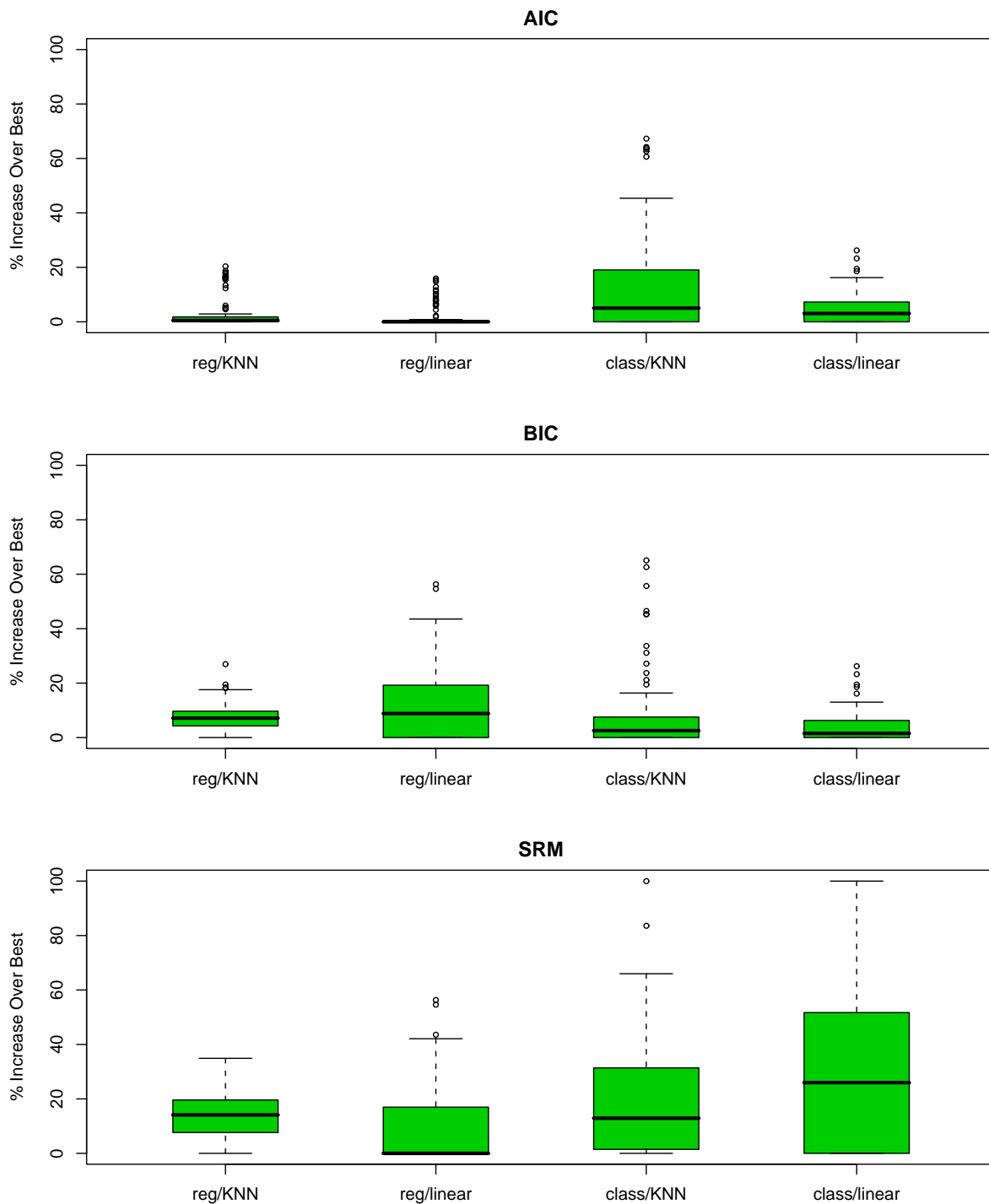
**FIGURE 7.4.** *AIC used for model selection for the phoneme recognition example of Section 5.2.3. The logistic regression coefficient function  $\beta(f) = \sum_{m=1}^M h_m(f)\theta_m$  is modeled as an expansion in  $M$  spline basis functions. In the left panel we see the AIC statistic used to estimate  $\text{Err}_{\text{in}}$  using log-likelihood loss. Included is an estimate of  $\text{Err}$  based on an independent test sample. It does well except for the extremely over-parametrized case ( $M = 256$  parameters for  $N = 1000$  observations). In the right panel the same is done for 0–1 loss. Although the AIC formula does not strictly apply here, it does a reasonable job in this case.*



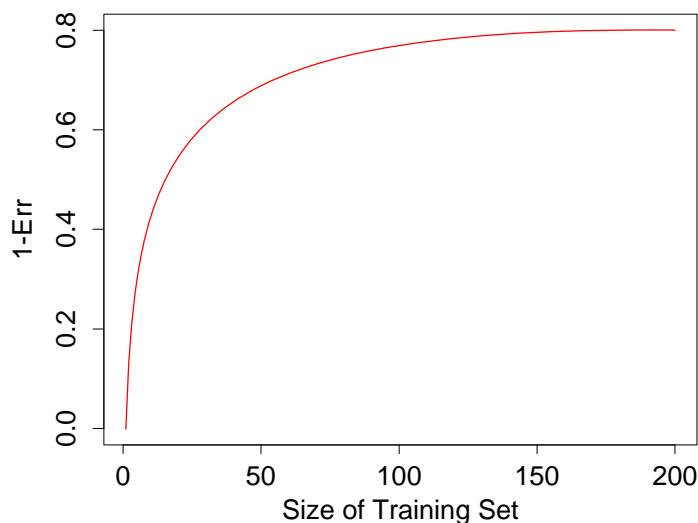
**FIGURE 7.5.** *The solid curve is the function  $\sin(50x)$  for  $x \in [0, 1]$ . The green (solid) and blue (hollow) points illustrate how the associated indicator function  $I(\sin(\alpha x) > 0)$  can shatter (separate) an arbitrarily large number of points by choosing an appropriately high frequency  $\alpha$ .*



**FIGURE 7.6.** *The first three panels show that the class of lines in the plane can shatter three points. The last panel shows that this class cannot shatter four points, as no line will put the hollow points on one side and the solid points on the other. Hence the VC dimension of the class of straight lines in the plane is three. Note that a class of nonlinear curves could shatter four points, and hence has VC dimension greater than three.*

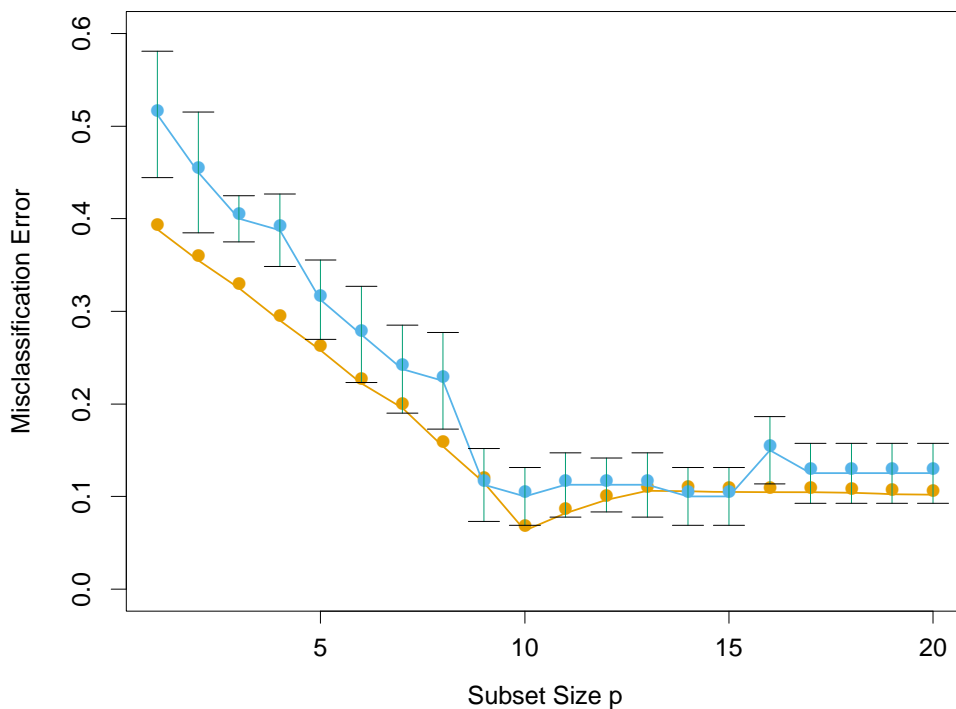


**FIGURE 7.7.** *Boxplots show the distribution of the relative error  $100 \times [\text{Err}_{\mathcal{T}}(\hat{\alpha}) - \min_{\alpha} \text{Err}_{\mathcal{T}}(\alpha)] / [\max_{\alpha} \text{Err}_{\mathcal{T}}(\alpha) - \min_{\alpha} \text{Err}_{\mathcal{T}}(\alpha)]$  over the four scenarios of Figure 7.3. This is the error in using the chosen model relative to the best model. There are 100 training sets each of size 80 represented*

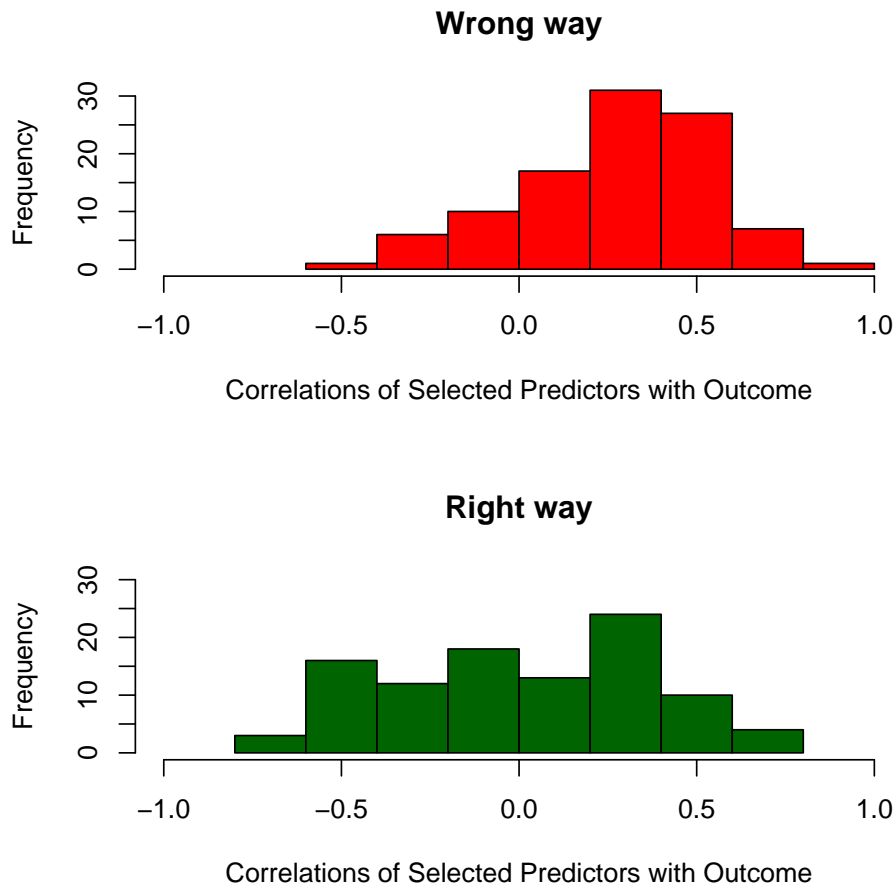


**FIGURE 7.8.** *Hypothetical learning curve for a classifier on a given task: a plot of  $1 - \text{Err}$  versus the size of the training set  $N$ . With a dataset of 200 observations, 5-fold cross-validation would use training sets of size 160, which would behave much like the full set. However, with a dataset of 50 observations fivefold cross-validation would use training sets of size 40, and this would result in a considerable overestimate of prediction error.*

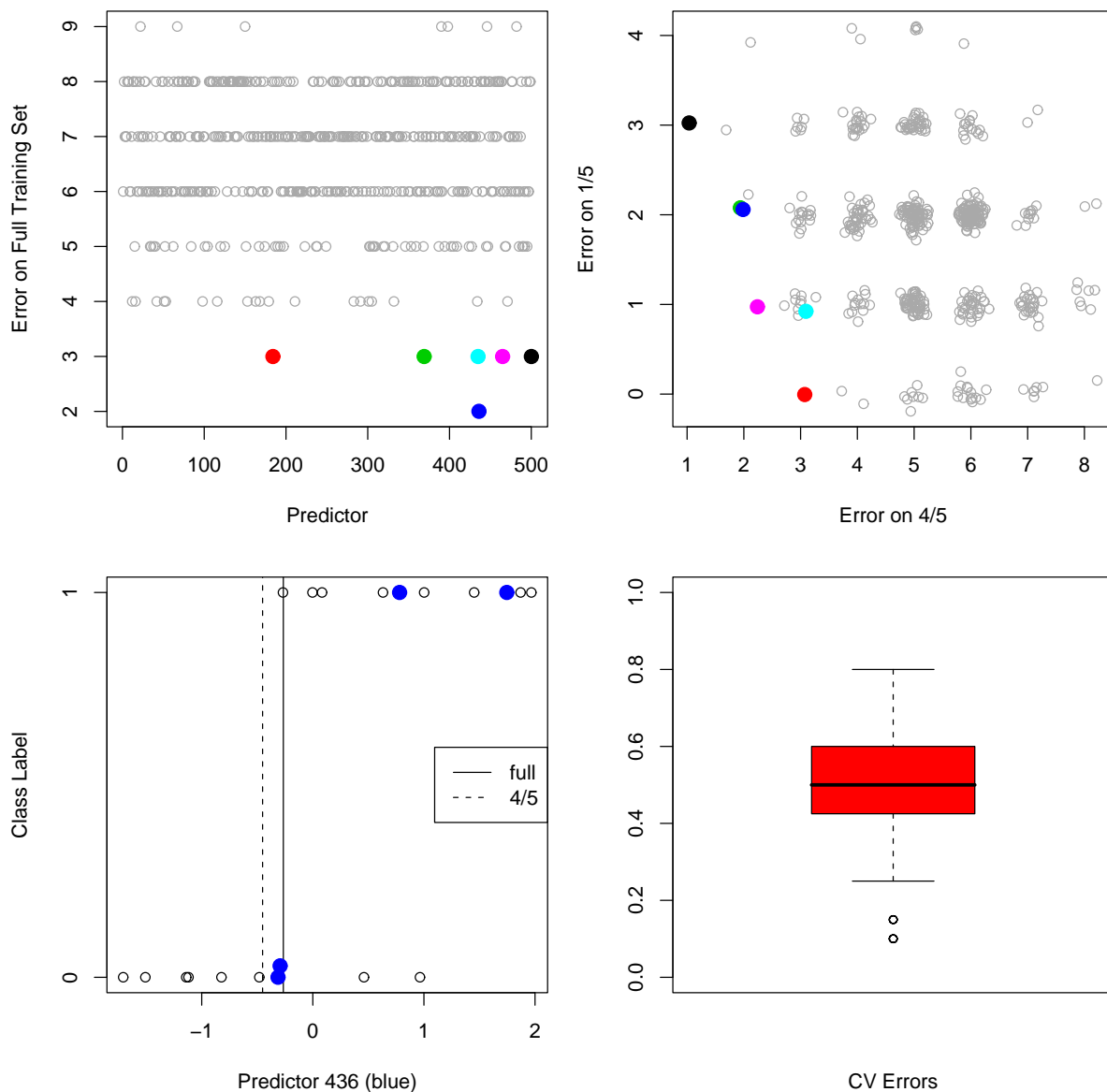




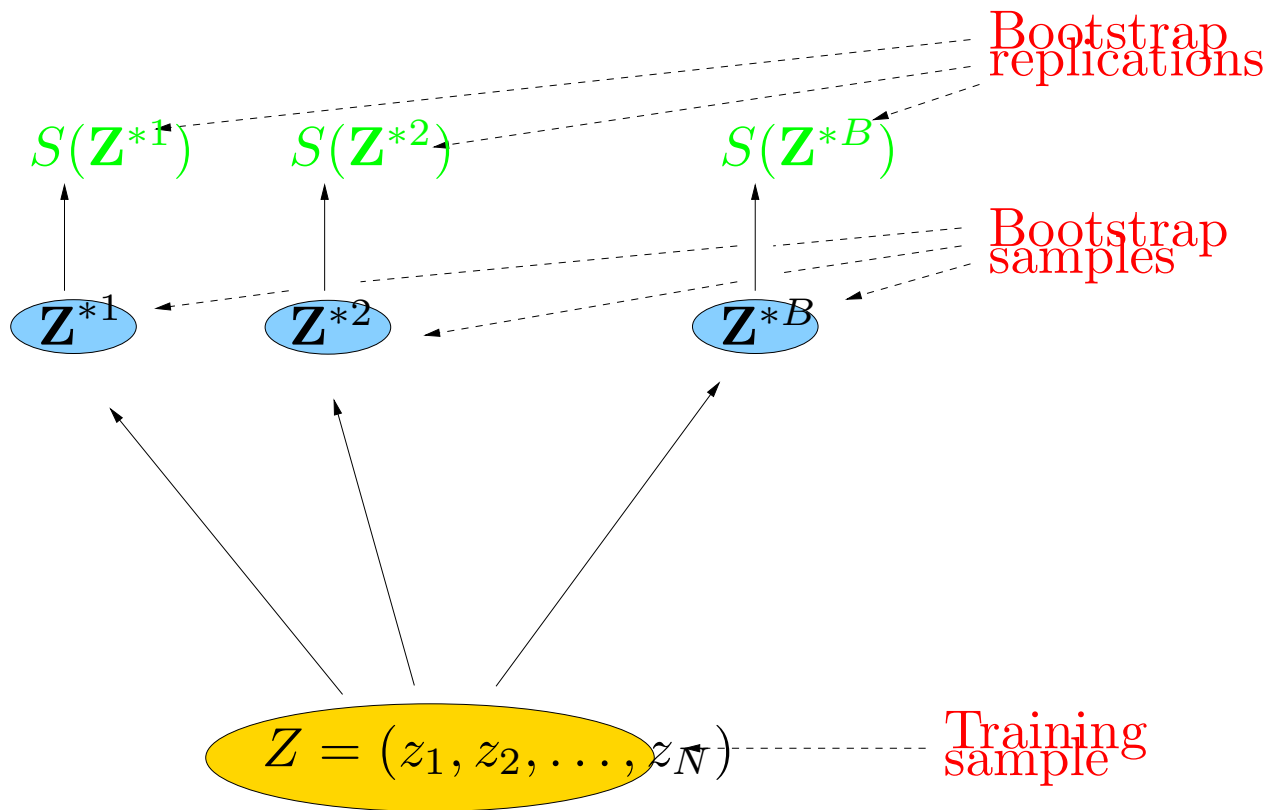
**FIGURE 7.9.** *Prediction error (orange) and tenfold cross-validation curve (blue) estimated from a single training set, from the scenario in the bottom right panel of Figure 7.3.*



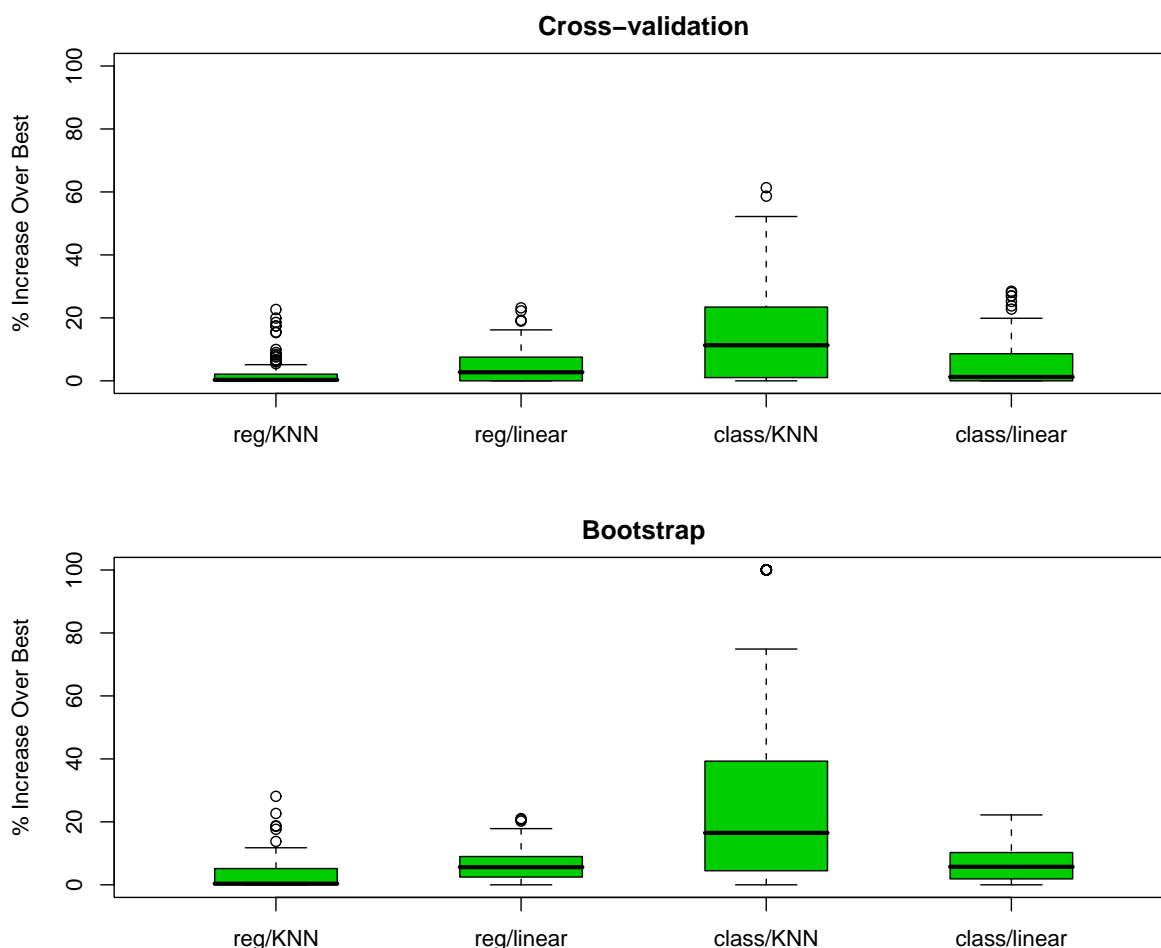
**FIGURE 7.10.** *Cross-validation the wrong and right way: histograms shows the correlation of class labels, in 10 randomly chosen samples, with the 100 predictors chosen using the incorrect (upper red) and correct (lower green) versions of cross-validation.*



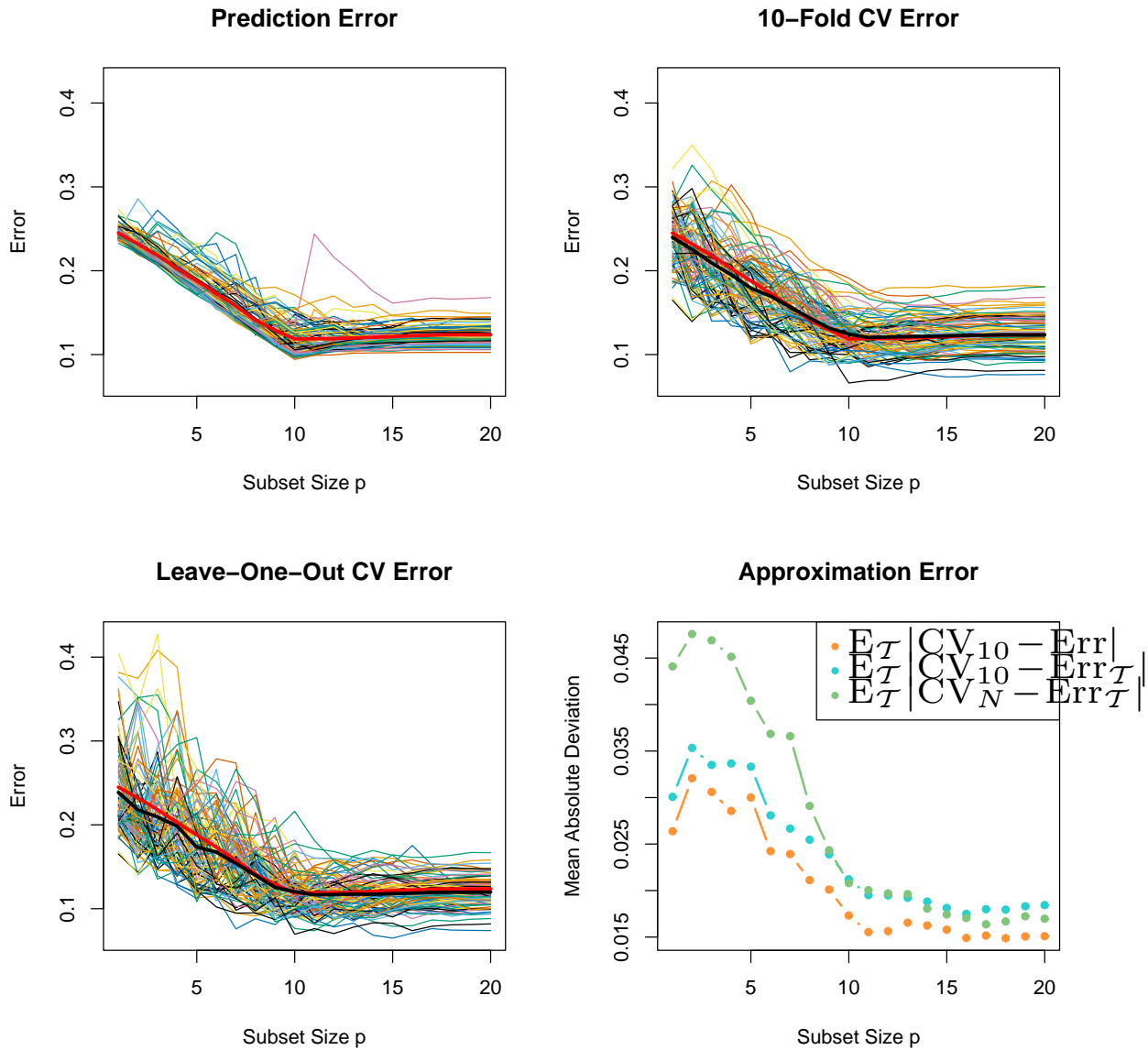
**FIGURE 7.11.** *Simulation study to investigate the performance of cross validation in a high-dimensional problem where the predictors are independent of the class labels. The top-left panel shows the number of errors made by individual stump classifiers on the full training set (20 observations). The top right panel shows the errors made by individual stumps trained on a random split of the dataset into 4/5ths (16 observations) and tested on the remaining 1/5th (4 observations). The best performers are depicted by colored dots in each panel. The bottom left panel shows the effect of*



**FIGURE 7.12.** Schematic of the bootstrap process. We wish to assess the statistical accuracy of a quantity  $S(\mathbf{Z})$  computed from our dataset.  $B$  training sets  $\mathbf{Z}^{*b}$ ,  $b = 1, \dots, B$  each of size  $N$  are drawn with replacement from the original dataset. The quantity of interest  $S(\mathbf{Z})$  is computed from each bootstrap training set, and the values  $S(\mathbf{Z}^{*1}), \dots, S(\mathbf{Z}^{*B})$  are used to assess the statistical accuracy of  $S(\mathbf{Z})$ .

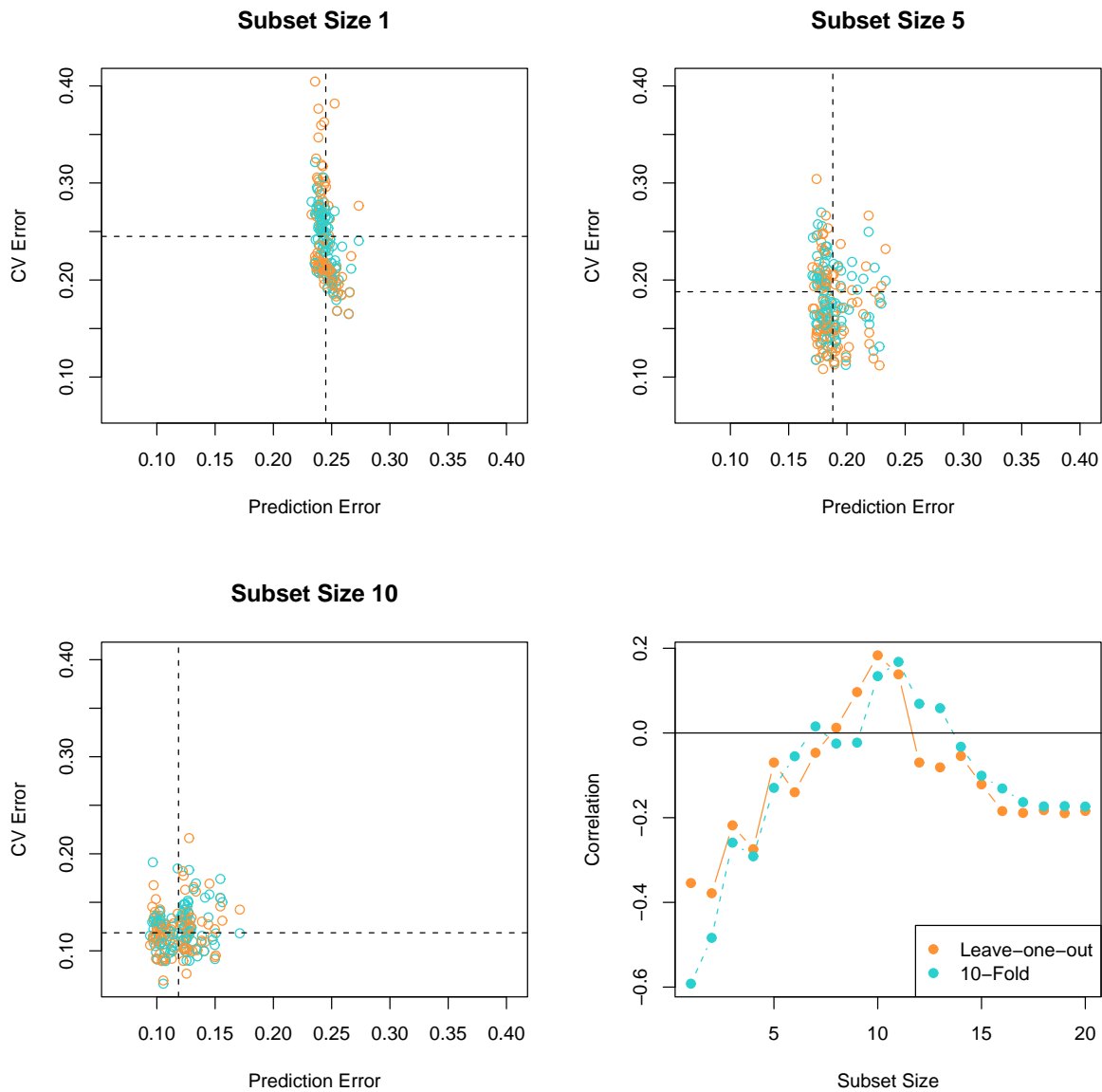


**FIGURE 7.13.** *Boxplots show the distribution of the relative error  $100 \cdot [\text{Err}_{\hat{\alpha}} - \min_{\alpha} \text{Err}(\alpha)] / [\max_{\alpha} \text{Err}(\alpha) - \min_{\alpha} \text{Err}(\alpha)]$  over the four scenarios of Figure 7.3. This is the error in using the chosen model relative to the best model. There are 100 training sets represented in each boxplot.*



**FIGURE 7.14.** Conditional prediction-error  $Err_{\mathcal{T}}$ , 10-fold cross-validation, and leave-one-out cross-validation curves for a 100 simulations from the top-right panel in Figure 7.3. The thick red curve is the expected prediction error  $Err$ , while the thick black curves are the expected CV curves  $E_{\mathcal{T}}CV_{10}$  and  $E_{\mathcal{T}}CV_N$ . The lower-right panel shows the mean absolute deviation of the CV curves from the conditional error,  $E_{\mathcal{T}}|CV_K - Err_{\mathcal{T}}|$  for  $K = 10$  (blue) and  $K = N$  (green), as well as from the expected error  $E_{\mathcal{T}}|CV_{10} - Err|$  (orange).





**FIGURE 7.15.** *Plots of the CV estimates of error versus the true conditional error for each of the 100 training sets, for the simulation setup in the top right panel Figure 7.3. Both 10-fold and leave-one-out CV are depicted in different colors. The first three panels correspond to different subset sizes  $p$ , and vertical and horizontal lines are drawn at  $\text{Err}(p)$ . Although there appears to be little correlation in these plots, we see in the lower right panel that for the most part the correlation is negative.*