SOLUTION for Homework 2 (10 pts total)

Problem 1

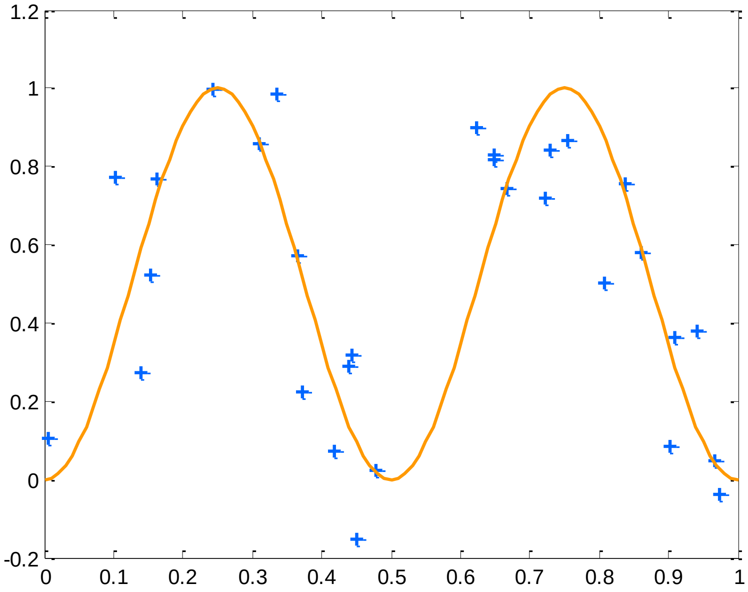


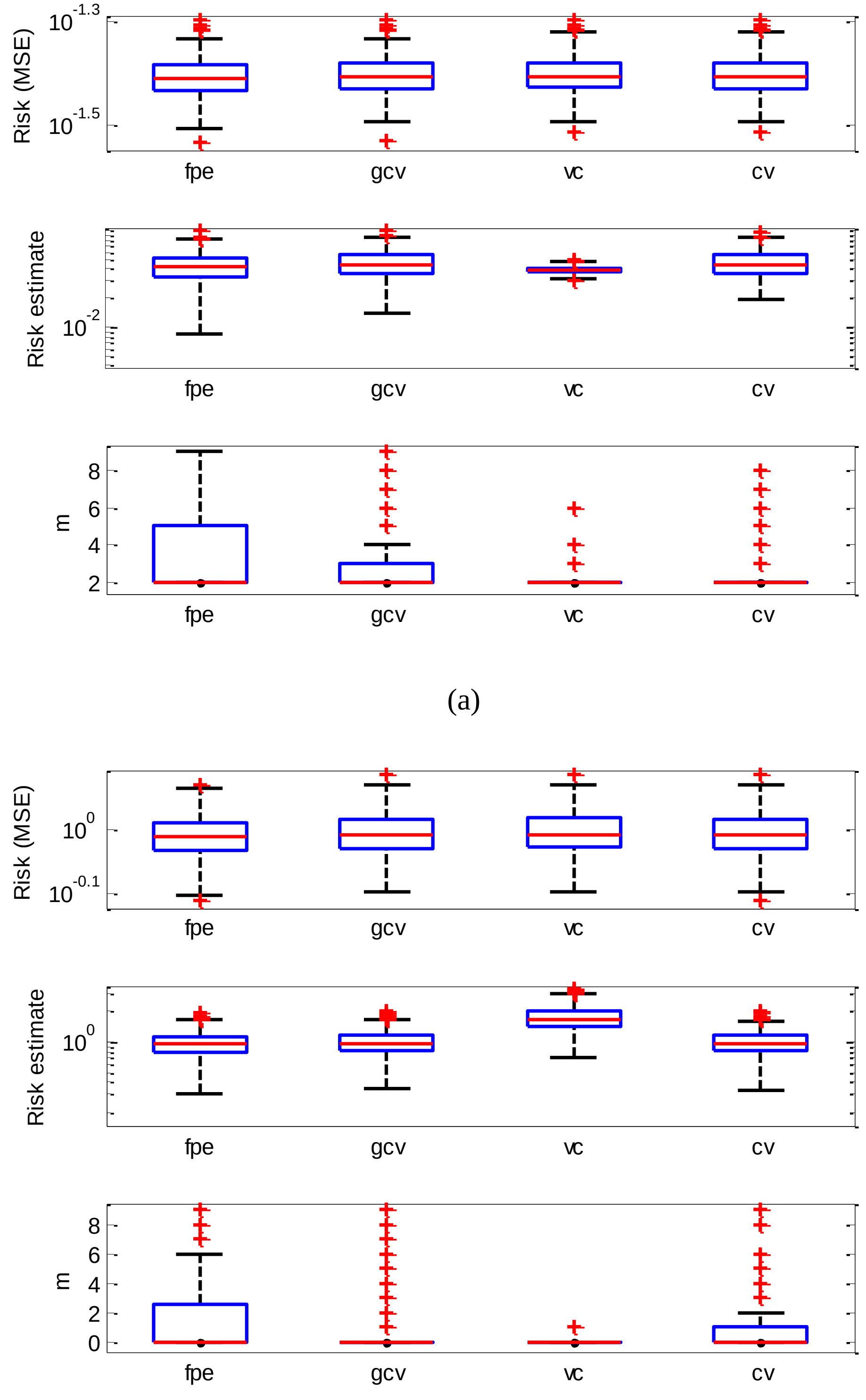
Figure 1. Training samples for one realization and the target function.

Note 1: For a given parameterization of approximating functions, i.e. trigonometric polynomials $f_m\left(x,\mathbf{w},\mathbf{v},b\right)=b+\sum_{i=1}^{m}w_i\sin\left(ix\right)+v_i\cos\left(ix\right)$, the VC-dimension is h=2m + 1.

Note 2: the target function sin^2 *(2π x)* = 0.5*(1- cos*(4π x)*). Hence, an optimal degree of the trigonometric polynomial should be chosen as m=2 (assuming we have enough training samples).

Empirical results below present 3 types of boxplots showing: model complexity (values of m) selected by each method, analytic prediction risk estimates (by each method), and actual prediction risk measured on independent test sets for sample sizes = 30 and 100.

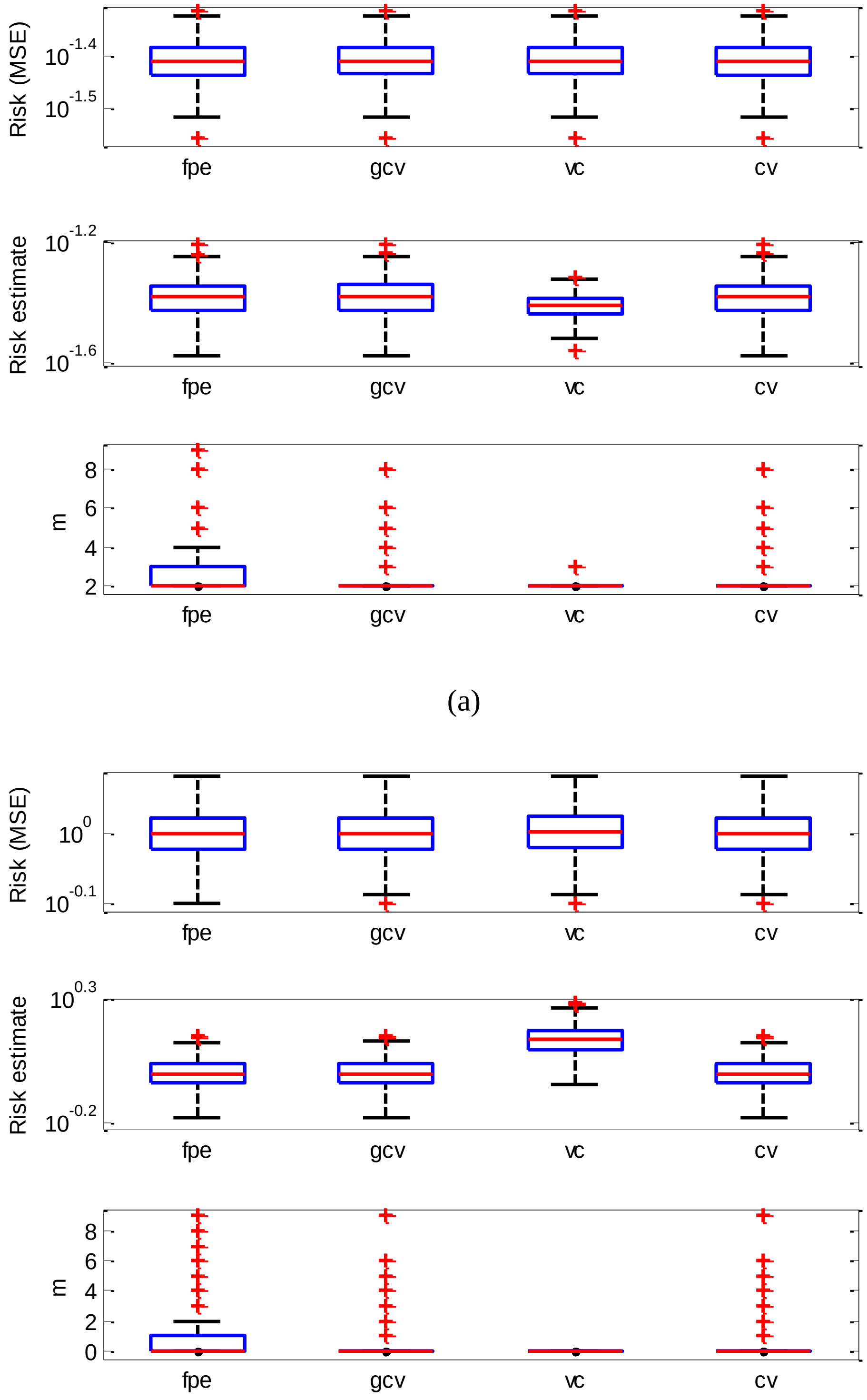
From the results, we can observe that the VC method typically tends to select the lowest DoF (comparing with other methods), and the MSE is comparable with others.



(a)

(b)

Figure 2. Model selection results for sample size 30 using trigonometric polynomials:
(a) sine-squared target function with additive noise $\sigma = 0.2$;
(b) Gaussian noise $\sigma = 1$.



(a)

(b)

Figure 3. Model selection results for sample size 100 using trigonometric polynomials:
(a) sine-squared target function with additive noise $\sigma = 0.2$;
(b) Gaussian noise $\sigma = 1$.

Problem 2 (Trading international mutual funds)

In this problem, we apply two learning methods, Linear Classifier and Quadratic Decision Boundary Classifier, to design a trading strategy for an international mutual fund (symbol **FIGRX**). Our learning model can be generally described as $y = f(x_1, x_2)$, where the input indicators, x_1 and x_2, are the daily closing prices percentage changes of the SP500 stock index (symbol **GSPC**) and the Euro-to-Dollar exchange rate (symbol **EURUSD**), respectively. The output y is the next-day **FIGRX** mutual fund value change in percentage.

Equivalently, the learning method will try to find a model f which gives good predictions on y. The trading decision rule is F = sign($\underline{y}$).

The classification results of a linear classifier are shown in Figure 4.
Figure 5 shows the cumulative account value for the trading strategy using this linear classifier.

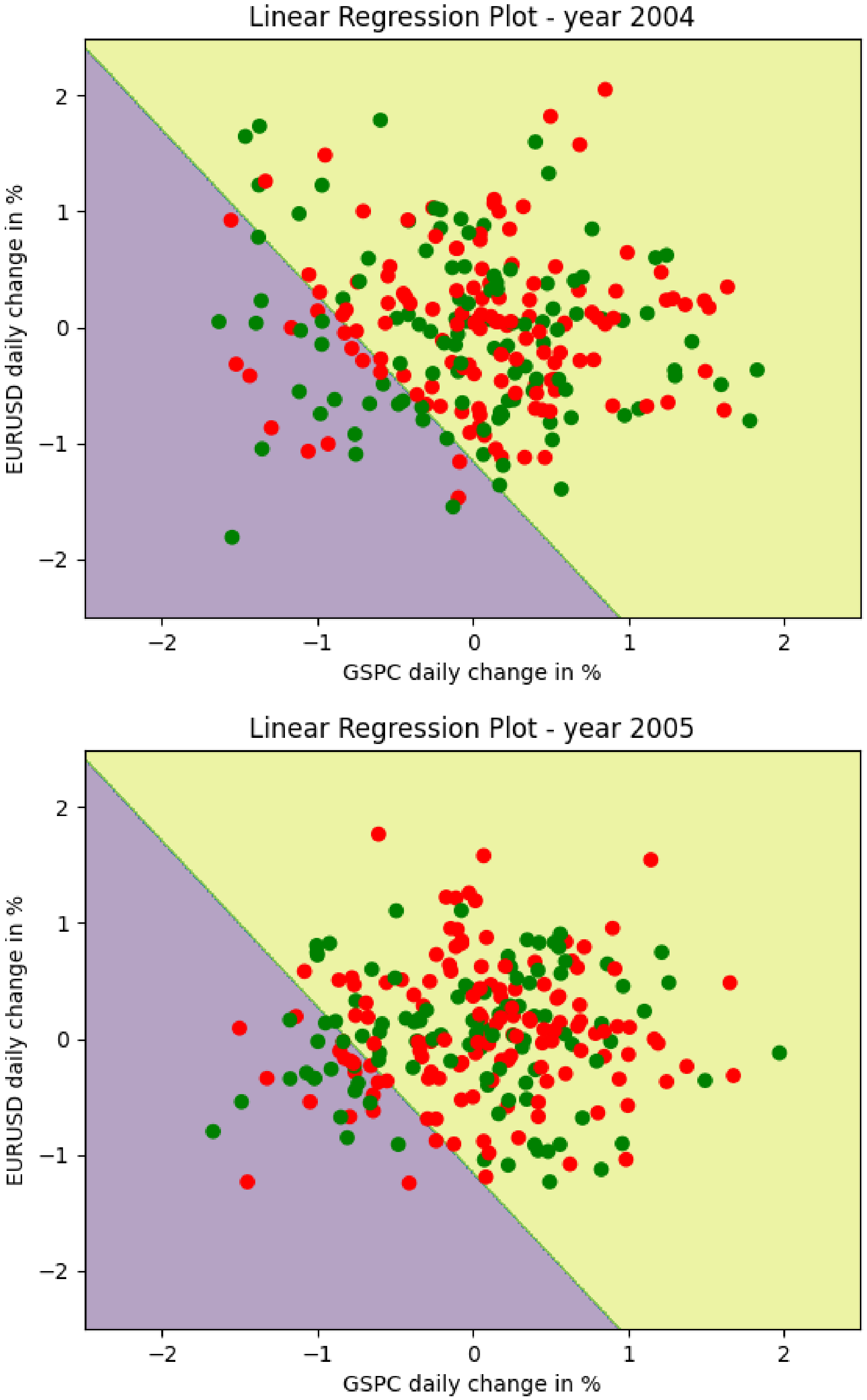


Figure 4. Linear decision boundary estimated using least squares regression overlapped on the training (2004) and test (2005) data.

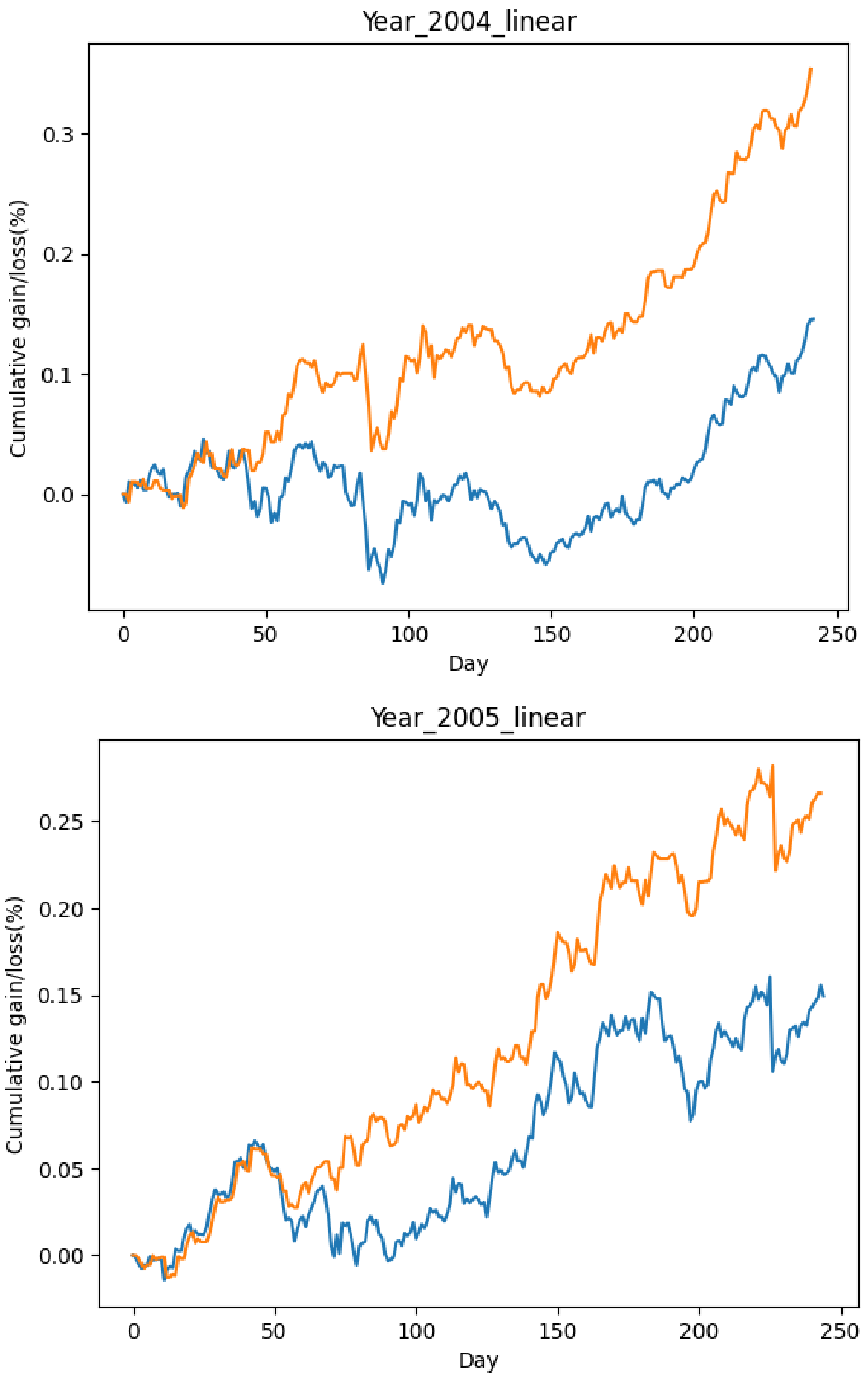


Figure 5. Effectiveness of the trading strategy using linear classifier. (a) Account gain/loss during training period (Year 2004). (b) Account gain/loss during test period (Year 2005).

The classification results of a quadratic decision boundary classifier are shown in Figure 6. Figure 7 shows the cumulative account value for the trading strategy using the quadratic decision boundary classifier.

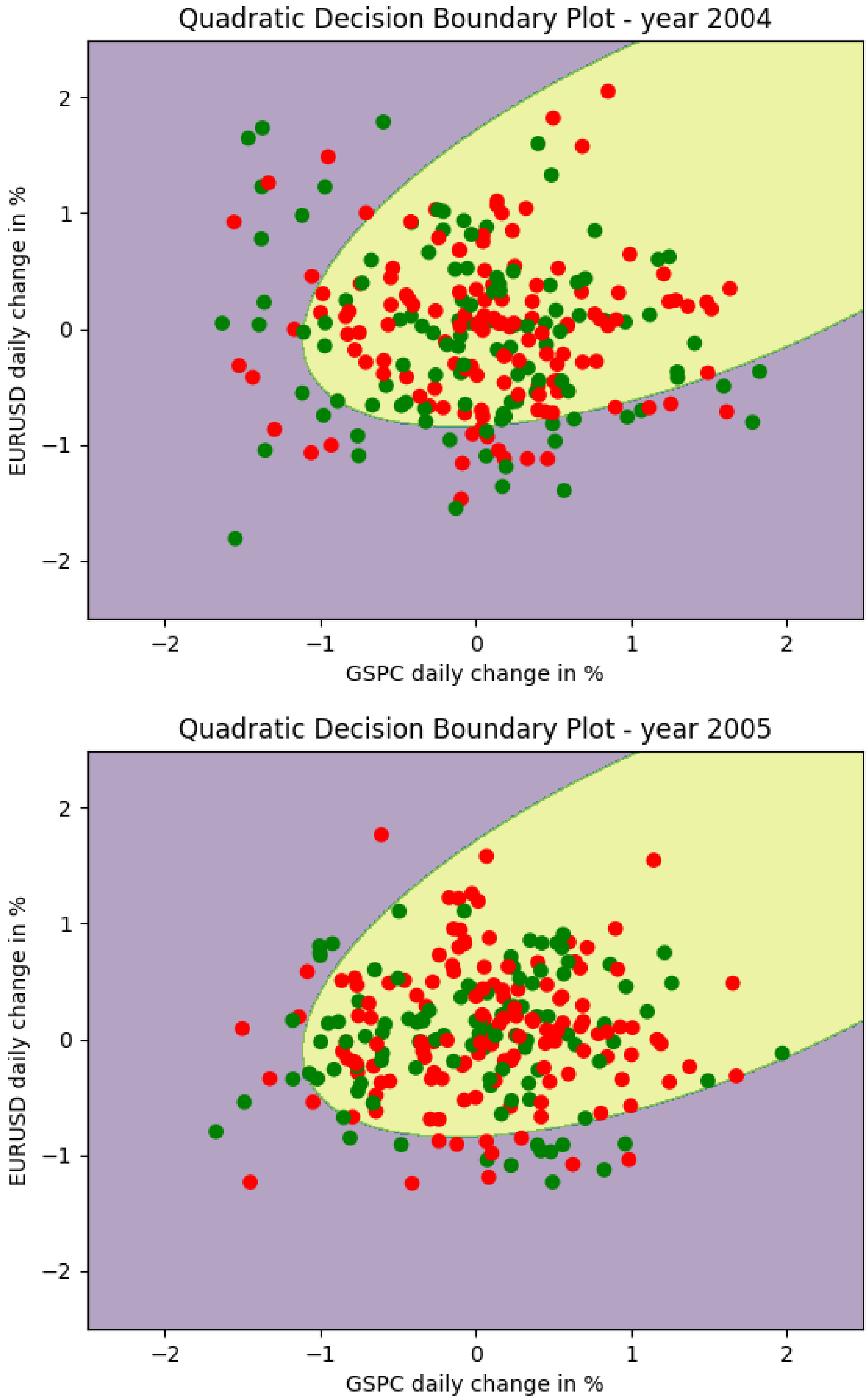


Figure 6. Quadratic decision boundary overlapped on the training (2004) and test (2005) data.

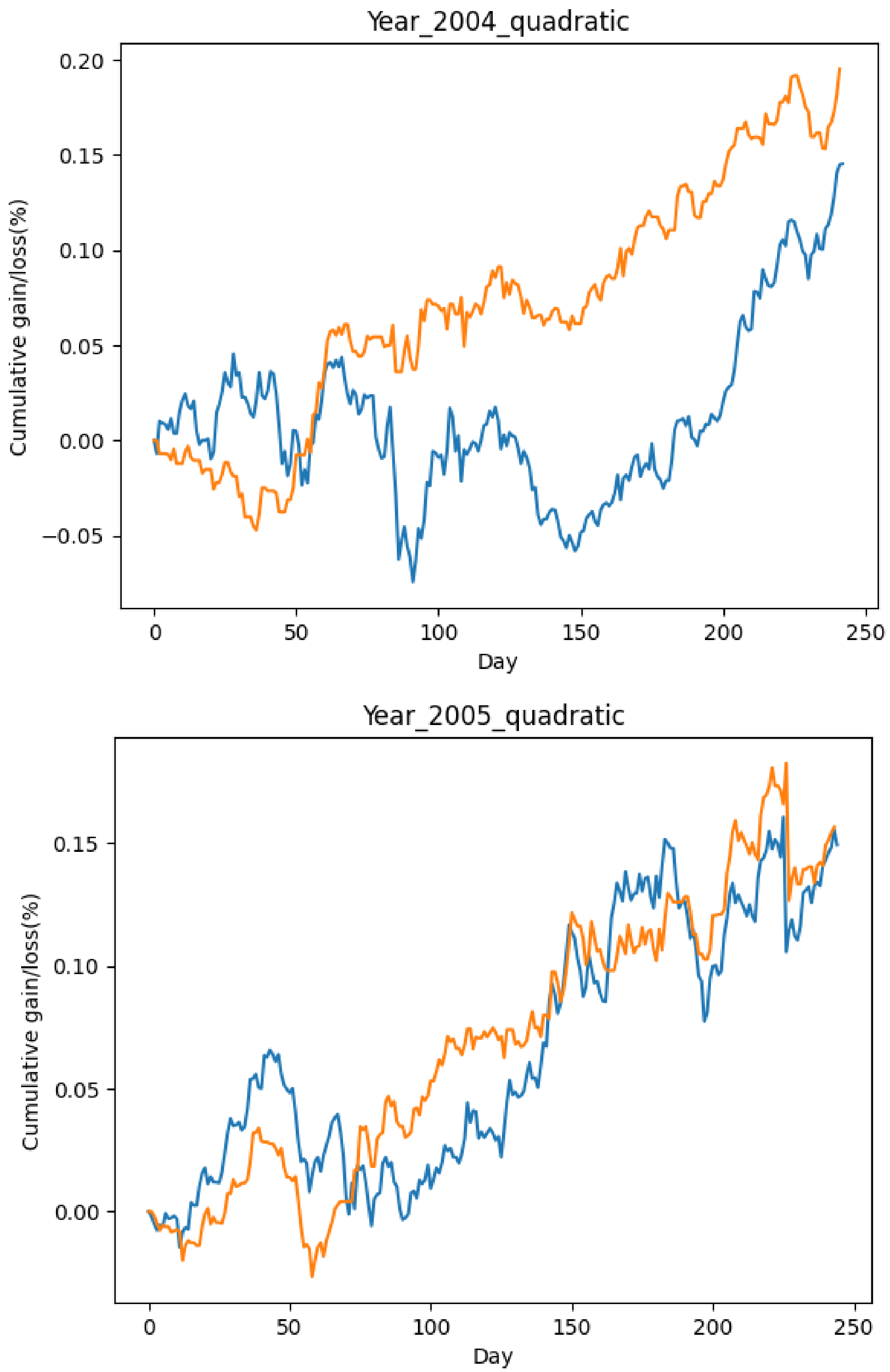


Figure 7. Effectiveness of the trading strategy using quadratic decision boundary classifier. (a) Account gain/loss during training period (Year 2004). (b) Account gain/loss during test period (Year 2005).

- Performance

  The training and test errors for two classifiers are summarized in Table 1. The errors are more than 40%. Note that the quadratic decision boundary classifier has the same training error as the linear classifier, but the test error is lower.

Table 1. The training and test errors in percentage.

	Training period 2004	Test period 2005
Linear	42.32	44.44
Quadratic	42.32	42.39

We compare the two classification approaches by the investment performances in Table 2. The *Gain* is the cumulative gain (or loss) in percentage at the end of the year. The Gain is calculated for linear and quadratic models and the *Buy-and-Hold* (B-H) strategy. Further, the market *Exposure* is the proportion of days when the account is fully invested in **FIGRX**. It also indicates how often a classifier assigns a sample to the positive (Up) class. Considering the gain, both methods have higher gain than the B-H method in both training and test periods. Comparing the two ML methods, the gain of the linear classifier during both periods is higher than that of the quadratic classifier. For the exposure, the quadratic decision boundary classifier achieves lower exposure in the training period but has higher exposure during the test period. So, arguably, the linear model is safer and more profitable.

Table 2. Summary of investment performance in percentage.

	Training period 2004			Test period 2005		
	Gain (ML)	Gain (B-H)	Exposure	Gain (ML)	B-H (B-H)	Exposure
Linear	35.37	14.54	85.06	26.60	14.91	84.77
Quadratic	19.54	14.54	76.76	15.63	14.91	85.19

- Discussion

  1. Can this trading strategy be used in practice?

     YES, based on empirical modeling results, trading strategy provides significantly better total return than ‘Buy-and-Hold’. The linear decision boundary classifier is better than the quadratic one because it achieves better performance than ‘Buy-and-Hold’ with lower market exposure (lower risk).

  2. Why learning methods do not use model complexity control?

     Because their model complexity (VC-dimension) is small relative to training sample size. So, good performance during training period implies good performance for test data. This follows from analysis of the VC generalization bound (for classification). Note that in this case, ‘performance’ is measured relative to ‘buy-and-hold’ strategy for the same security, rather than absolute returns during the test period.