Machine Learning and Market Indices

An Exercise In Modelling Stock Market Data

Trevor H. Drees

Contents

Disclaimer **2**

Introduction **2**

Data **2**

Analyses **4**

Overview of Methods 4

Logistic Regression 5

Discriminant Analysis 9

Support Vector Machines 12

Decision Trees 12

Advanced Methods 12

Additional Trading Strategies 12

Conclusions **12**

References **12**

Figures **12**

Disclaimer

Trading stocks, ETFs, options, futures, or any other financial assets or derivatives has inherent risks that should be understood before making any substantial investments; such risks can include significant loss of principal, and it is the investor's responsibility to determine their appropriate level of exposure to risk. The contents of this work are not, in any form, meant to serve as investment advice, but rather as exploratory analysis of stock market data with the goal being able to predict changes in future conditions based solely off of conditions at a prior date. As such, the author bears no liability for any damages or losses, realised or unrealised, incurred by the reader should they choose to use the models in these analyses for any sort of investment guidance.

Introduction

Stock market data can be incredibly difficult to predict from just stock prices alone; changes in value of a particular stock or ETF today, last week, or last month do not necessarily dictate how it will move today. Often, fluctuations in value are instead driven by shifts in investor confidence due to changes in government monetary policy, outlook on a company's so-called "fundamentals", or disruptions in activity of a particular company or sector, among many other possible factors. While a well-diversified portfolio will almost certainly increase in value over the long run without frequent intervention, the ability to predict short-term changes and fluctuations based off market data is incredibly useful to active traders that are less concerned with holding long positions. For investors that hold particular investments over the course of only days or weeks, short-term changes are much more important for deciding when to buy and sell a stock or whether to make a put or call when trading options.

Here, we seek to create models that can predict short-term market behaviour, as such a model would be a very valuable investment tool for traders. More specifically, we are interested in predicting whether the market will increase or decrease in a given week based on the performance of the previous 5 weeks (25 trading days). To build and test our models, we use data from the S&P 500, a major U.S. stock market indices; we also include data from the NASDAQ Composite and Dow Jones Industrial Average and though they are not part of our analyses, the code we use can easily be adapted to predict increases and decreases on these other two indices. For the S&P 500 data, we compare models derived from a variety of classification frameworks, including methods such as logistic regression, linear and quadratic discriminant analysis, support vector machines with a variety of kernel types, and decision trees with bagging and random forests. We also discuss possible trading strategies based on model attributes such as overall accuracy, sensitivity, specificity, and positive and negative predictive value.

Data

We first start by introducing and briefly exploring the six data sets that will be used in these analyses. First, we have `SPDaily` and `SPWeekly`, which are daily and weekly data (respectively) on the S&P 500, a major stock market index consisting of 500 publicly-traded American companies. Next, we have `NDDaily` and `NDWeekly`, which are daily and weekly data (respectively) on the NASDAQ Composite, another major stock market index that is largely weighted on the information technology sector. Finally, we have `DJDaily` and `DJWeekly`, which are daily and weekly data (respectively) on the NASDAQ Composite, a major stock market index that consisting of 30 large American companies from a variety of different market sectors. All data, which are available online from Yahoo Finance[1], were scraped in Python using the `yfinance` packeage and stored in six different CSV files; weekly/daily percent change was then calculated using the opening

[1] S&P 500 data can be found here, NASDAQ Composite data can be found here, and Dow Jones Industrial Data can be found here. Data last accessed on 26 March 2021.

and closing prices for a particular week/day, and lagged up to 5 weeks/days. The total and average percent change over a 5-week or 5-day period were also calculated. The variables in our new datasets are listed below:

- **BVolume**: average number of daily shares traded (billions)
- **PctChange**: percent change for a given week/day
- **Direction**: whether the percent change for a given week/day was negative or positive
- **Prev5GM**: Average (geometric mean) percent increase over the previous 5 weeks/days
- **Prev5Pct**: Total percent increase over the previous 5 weeks/days
- **Lag1**: 1-week lag on percent change for a given week/day
- **Lag2**: 2-week lag on percent change for a given week/day
- **Lag3**: 3-week lag on percent change for a given week/day
- **Lag4**: 4-week lag on percent change for a given week/day
- **Lag5**: 5-week lag on percent change for a given week/day
- **VLag1**: 1-week lag on volume for a given week/day
- **VLag2**: 2-week lag on volume for a given week/day
- **VLag3**: 3-week lag on volume for a given week/day
- **VLag4**: 4-week lag on volume for a given week/day
- **VLag5**: 5-week lag on volume for a given week/day

Note that for the `SPDaily`, `NDDaily`, and `DJDaily` datasets, all calculations and quantities are in terms of days rather than weeks; for the `SPWeekly`, `NDWeekly`, and `DJWeekly` datasets, all calculations and quantities are in terms of weeks rather than days. For example, the `Lag2` variable would represent the percent change two days ago in the `SPDaily` dataset, but would represent the percent change two weeks ago in the `SPWeekly` dataset.

We first conduct some exploratory analyses of the data to see if there are any patterns that need attention or may be important for our analyses, starting with the S&P 500 daily and weekly data. Upon examining the correlation matrices and heatmaps shown in Figure 1, there seems to be several strong correlations. The strong positive correlation between `Direction` and `PctChange` is not surprising, given that `Direction` is literally defined by whether the percent change is negative or positive. Though we would thus expect there to be a perfect relationship between the two variables, and indeed there is, it does not translate well to a linear correlation. We also see that `Prev5GM` and `Prev5Pct` have a correlation of 1; this is because the total percent increase over the previous 5 weeks can also be found by exponentiating the average percent increase to the fifth power. Thus, there is an obvious relationship between the two variables, and any models containing the both may experience multicollinearity issues. These same trends can also be seen in the NASDAQ data (Figure 2) and the Dow Jones data (Figure 3).

Another thing worth pointing out in Figure 1 is that `Direction`, the variable that we're interested in predicting, bears almost no correlation with `Prev5GM`, `Prev5Pct`, or any of the `Lag` variables; this may be problematic when trying to make predictions. Figure 4 provides further evidence of this on the S&P 500 weekly data and Figure 5 on the S&P 500 daily data, focusing specifically on the lack of correlation between percent change and the five lag variables. Here, we see that there is almost no autocorrelation of percent change in a given week; that is, percent change in a given week is not strongly correlated with percent change in any of the five weeks before. This observation may be bad news for our analyses later: if there is not much of an association between percent change in a given week and percent change in any of the weeks before it, then how can we use past weekly performance to accurately predict future weekly performance? While there is no apparent autocorrelation in returns for the S&P 500, there does seem to be an autocorrelation in trading volume, as is evident in Figures 6 and 7.

Unsurprisingly, we also see the exact same lack of autocorrelation in returns and strong autocorrelation in trading volume for both the NASDAQ and Dow Jones data. Figures 8 and 9 demonstrate a lack of autocorrelation in NASDAQ daily and weekly returns, while Figures 10 and 11 show that daily and weekly volume is strongly autocorrelated; likewise, Figures 12 and 13 also demonstrate a lack of autocorrelation in Dow Jones daily and weekly returns, while Figures 14 and 15 again show that daily and weekly volume is strongly autocorrelated.

Separating the S&P 500 data based on the value of `Direction` allows us to further explore trends in increases and decreases. For example, upon doing so, we see that the mean weekly percent decrease on the S&P 500 is -1.81%, while the mean weekly increase is 1.67%; however, even though the average magnitude of weekly percent decreases is higher than that of weekly percent increases, the number of weeks in which there was an increase is greater than the number of weeks in which there was a decrease (1041 versus 784). If we look at the data in terms of days instead, we find that there are again more increases than decreases, with 4063 negative days with a mean daily decrease of -0.752% and 4750 positive days with a mean daily increase of 0.707%; this difference is noticable in the bottom panel of Figure 16. The fact that there are more positive than negative days/weeks is likely responsible for the trend shown in the top panel of Figure 16 where, despite plummeting several times (e.g. two recessions in the 2000's and the COVID-related plunge in early 2020) and suffering numerous smaller drops, the S&P 500 was still almost 14 times as high in 2021 as it was in 1986.

Again, we see the same trends in the NASDAQ and Dow Jones data if we separate it based on `Direction`. The NASDAQ has 772 negative weeks with a mean weekly decrease of -2.36% and 1052 positive weeks with a mean weekly increase of 2.04%; when broken into days, there 4000 negative days with a mean daily decrease of -0.849% and 4776 positive days with a mean daily increase of 0.724%. This difference is noticable in the bottom panel of Figure 17 and is likely responsible for the increase in value over time, despite the massive crash in 2000 from the Dot-Com Bubble. If we look at the Dow Jones, we see that it has 643 negative weeks with a mean weekly decrease of -1.64% and 866 positive weeks with a mean weekly increase of 1.53%; when broken into days, there 3383 negative days with a mean daily decrease of -0.728% and 3913 positive days with a mean daily increase of 0.696%. This difference is again noticable in the bottom panel of Figure 18 and is also likely responsible for the increase in value over time. Note that because the Yahoo Finance data on the Dow Jones only starts in 1992 (though the Dow Jones has been around much longer than that), we have less data to work with than we do for the S&P 500 and NASDAQ.

We now move forward with our analyses, keeping in mind that the lack of autocorrelation in weekly percent change may make it difficult to accurately construct models that use past percentage changes to predict future index movements. However, there is some good news: because there is only a slight difference in the number of positive and negative days/weeks, we do not have to worry about predicting a small number of observations in a highly-skewed data set, and thus face fewer restrictions with the tools and algorithms we plan on using. Before we begin, we split each of our data sets into a training set with observations from 1986-2010 and a test set with the remaining observations from 2011-2021; this validation set approach will allows us to fit models on the training data and then examine their performance on the test data, reducing the chances of overfitting the data. We also drop the variables `VLag1` through `VLag5`, as preliminary analyses (not shown in this report) indicate that these variables make almost no difference when constructing our models.

For the sake of brevity, we will focus on analysing the S&P 500 data, though the code for our analyses can easily be extended to accommodate the NASDAQ and Dow Jones data.

Analyses

Overview of Methods

First, we begin our analyses by using a binary-response logistic regression on the data to predict whether a given trading session will see an increase or decrease in value on the S&P 500, fitting all possible combinations of non-interactive predictors and then selecting top models based on overall accuracy and positive predictive value (PPV); we then examine a much larger pool of models that include interactive terms using backward and forward step algorithms to maximise accuracy and PPV. Second, we perform similar analyses on non-interactive and interactive discriminant analysis models, also comparing accuracy and PPV between linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA), which are two of the most commonly-used types of discriminant analysis. Third, we examine the performance of non-interactive and interactive models by maximising accuracy and PPV on support vector machines with linear and radial kernels.

Finally, we use decision tree models to again maximise accuracy and PPV on interactive and non-interactive models, and then expand our analyses to include bagging and random forests.

Logistic Regression

We begin our analyses on the S&P 500 data by using a logistic model to predict increases or decreases in the S&P 500, starting with the weekly data before proceeding to the daily data. We first consider a model such that `Direction` is the response and `BVolume`, `Prev5GM`, `Prev5Pct`, and`Lag1` through `Lag5` are predictors, and fitting this model to the training data.

```r
glm(Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 + BVolume + Prev5GM + Prev5Pct,
    data = SPWeeklyTrain, family = binomial)
```

This model contains the maximum number (8) of non-interactive predictors and is only one of many models that we can make; we could fit a model with only seven of these eight terms, six of these eight terms, five of these eight terms, and so on. As a matter of fact, one can prove that the total number of non-interactive models that we can form is

$$n = \sum_{i=1}^{8} \frac{8!}{i!(8-i)!} = 255 \tag{1}$$

which is computationally feasible if we wish to evaluate every single one of these possible models. However, fitting these models individually like we did in the code above would be too time consuming, so we write a function `logreg.comb` to do it for us.

```r
logreg.comb <- function(dataTrain, output, n){
  combn(names(dataTrain)[-c(2:3)], n) %>%
    apply(FUN = paste0, MARGIN = 2, collapse = "+") %>%
    paste0("Direction~", .) %>%
    sapply(FUN = logreg, dataTrain = dataTrain, output = output) %>%
    sort() %>%
    return()}
```

Here, we use the `combn` function to generate all 255 different combination of variables and then `paste0` to collapse them down into their individual formulas. We then use `sapply` to apply a custom function `logreg` to each of these formulas; this custom function fits a logistic regression using a given formula and training data set and returns a specified statistic such as accuracy, sensitivity, specificity, positive predictive value, or negative predictive value. Because this function is too long to properly include in a code snippet, we have not included it here and can instead be found in the markdown for this analysis.

After creating the functions `logreg` and `logreg.comb`, we then use them to evaluate all 255 possible non-interactive models and find the one that maximises training set accuracy. Doing so yields the model

$$logit(\pi) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 \tag{2}$$

with the predictor variables `BVolume` `Prev5GM`, `Lag2`, `Lag3`, and `Lag5`, and response π representing the probability of an increase in value. We will call this model **Model 1**. This model has an accuracy of approximately 57.69% on the training data which, while low, still seems to be slightly better than guessing. However, we see that this accuracy is highly skewed towards predicting increases; our model has a training sensitivity of 91.72% and correctly classified 665 out of 725 increases, but has an extremely low specificity of 14.78% and correctly classified only 85 out of 575 decreases. Sensitivity and specificity may not be useful in the context of this problem, though; recall that they would interpreted such that given an increase (or decrease), the sensitivity (or specificity) is the probability of correctly identifying that increase (or decrease). However, if we are trying to predict the behaviour of a stock or ETF next week, then we are not given an increase or decrease then since that is what we are trying to predict in the first place! As such, it might make more sense to examine something like positive predictive value (PPV); that is, when the model predicts an increase, the percentage of times that an increase actually occurs. If we use `logreg.comb` again, but this time seeking to maximise the training PPV, we actually get the same model, with a PPV of 57.58% on the training data. We will call this model **Model 2** to distinguish it from Model 1 because though they are the same model, they were selected using different criteria. We can also try and maximise negative predictive value (NPV); that is, when the model predicts a decrease, the percentage of times that a decrease actually occurs. Upon doing so, we get the model

$$logit(\pi) = \beta_0 + \beta_1 x_1 \tag{3}$$

with `Prev5GM` as the only predictor. This model has an NPV of 100%, meaning that every time a decrease is predicted, the markets actually decrease. This number is suspiciously high, and after checking the confusion matrix, it is clear why: the model only predicted a decrease once and got it right, while the other 542 decreases were incorrectly predicted as increases. Even looking at the next few highest models in terms of NPV, we see the same trend of high NPVs due to a extremely low number of decreases that just so happened to be correctly predicted. As such, it might be wise to stick with accuracy and PPV when choosing our models.

We can also examine the training accuracy of models that include two-way interactions, as these interactions may help capture patterns in the data that our original non-interactive models might have missed. The total number of two-way interaction terms plus single terms is 36, so thus the total number of interactive models that we can form is

$$n = \sum_{i=1}^{36} \frac{36!}{i!(36-i)!} = 68719476735. \tag{4}$$

which is much larger than the number of possible models considering only non-interactive terms. Evaluating every single one of these possible models would be far too computationally expensive for R, and we thus cannot calculate the accuracy of each model like we did when we only had 255 models. However, we can develop a heuristic solution that only searches through several of these possible models rather than attempting to test them all; here, we create an algorithm that performs a backward search much like how the function `step` with argument `direction = "backward"` does, though selecting our models based on training data accuracy rather than Akaike's information criterion (AIC).

```r
# Function to remove a term and find new accuracy
logreg.lcv1 <- function(i, dataTrain, output){
  predVec[-i] %>%
    paste0(., collapse = "+") %>%
    paste0("Direction~", .) %>%
    logreg(dataTrain = SPWeeklyTrain, output = output) %>%
    return()}

# Function to return accuracy and new list of predictors after removing a predictor
logreg.bwd <- function(predVec, output){
  accs <- sapply(1:length(predVec), FUN = logreg.lcv1, dataTrain = SPWeeklyTrain,
                 output = output)
  pv <- predVec[-which.max(accs)]
  return(list(predVec = pv, maxacc = max(accs)))}
```

Here, we start with the maximum number (36) of interactive and non-interactive terms, and use an `output` of `accuracy` to select our models based on overall prediction accuracy. We then calculate the new training accuracy after we remove a term, and do this for each term, then removing the term that results in the highest accuracy. We repeat this process until the model accuracy is maximised. Doing so results in a model with a training accuracy of 59.77% and 32 terms; we will call this model **Model 3**. Model 3 has the structure

$$logit(\pi) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_{32} x_{32} \tag{5}$$

and has so many terms that it would not be practical to list them here. We can also perform forward selection, starting with a null model and calculating the new training accuracy after adding a term, doing this for each term, and then adding the term that results in the highest accuracy.

```r
# Function to add a term and find new accuracy
logreg.lcv2 <- function(i, dataTrain, output){
  c(predVecFwd, predVecRemain[i]) %>%
  paste0(., collapse = "+") %>%
    paste0("Direction~", .) %>%
    logreg(dataTrain = SPWeeklyTrain, output = output) %>%
    return()}

# Function to return accuracy and new list of predictors after adding a predictor
logreg.fwd <- function(predVecRemain, predVecFwd, output){
  accs <- sapply(1:length(predVecRemain), FUN = logreg.lcv2, dataTrain = SPWeeklyTrain,
                 output = output)
  pvf <- c(predVecFwd, predVecRemain[which.max(accs)])
  pvr <- predVecRemain[-which.max(accs)]
  return(list(predVecFwd = pvf, predVecRemain = pvr, maxacc = max(accs)))}
```

Doing so now results in a model with a training accuracy of 57.15% and 7 terms; we will call this model **Model 4**. While the training accuracy is lower on this model, there are fewer terms than our Model 3 resulting from backward selection. We can also use the same forward and backward algorithms but select models using PPV instead of accuracy. If we do so using the backward step algorithm, we get a model with a PPV of 59.24% and 32 terms; we will call this model **Model 5**. While this model has the same number of terms as Model 3, the terms and coefficient estimates are not identical. Using the forward step algorithm, we get a model with a PPV of 57.71% and 10 terms; we will call this model **Model 6**.

Now that we have three models (1, 3, and 4) selected by maximising overall accuracy and three models (2, 5, and 6) selected by maximising PPV, we can compare their performance. Model 3 has the highest overall accuracy on the training data at 59.77% and, interestingly, also has the highest PPV at 59.32%; the fact that this model has a higher PPV than any of the models that were explicitly selected by maximising PPV suggests that the heuristic algorithm used in Model 5 and Model 6 may have only found a local maximum rather than a global maximum. We can also plot an ROC curve for each model, as has been done in Figures 19 and 20; here, we see that Model 3 has the second-highest area under the curve (AUC) at 0.5893. Model 5 has the highest AUC at 0.5900, though only by a slight margin, and has only a slightly lower accuracy (59.62%) and PPV (59.24%) than Model 3, making these two models extremely competitive.

However, we also need to assess model performance on the test data. We again see that model 3 has the highest accuracy (64.38%) and highest PPV (64.18%), and this time even has a higher AUC (0.6538) than Model 5. Model 5 is still a close second, though, with an accuracy of 64.00%, PPV of 63.96%, and AUC of 0.6428.

We can also check the residuals for these models to ensure that the logistic model is an appropriate fit. As can be seen in Figures 21 and 22, most of the residuals are resemble those of a typical logistic regression: two bands above and below zero, with a loess fit that approximates a horizontal line at zero. However, significant curvature of the loess fit to the residuals of Model 4 and Model 6 suggest that the logistic model may not be satisfactory in these instances.

So far, we see that Model 3 and Model 5 have overall accuracies and PPVs close to 60%, which is surprisingly high considering the lack of autocorrelation in our data and how noisy stock market data can be. But how can we connect model accuracy and PPV back to return on investment? While accuracy and PPV are excellent ways to understand the predictive capabilities of our models, we also seek to understand whether or not these models can "beat the market", so to speak.

We can do this by applying the models to the test data, and using model predictions to calculate the cumulative percent change at each point in time, as has been done in the left panels of Figures 23 and 24. The black lines represent a buy-and-hold strategy without any action informaed from the model; that is, an investor would simply buy shares and hold them without selling, regardless of the increases and decreases in value along the way. We see that under a buy-and-hold strategy, an investor that purchased shares at the beginning of 2011 would find that their investment would be worth approximately 2.30 times as much at the beginning of 2021, which represents an approximately 130% increase in value over the course of a decade. The blue lines represent the strategy of selling all shares before a predicted decrease, then buying them back before the next predicted increase, effectively sheltering all assets from a suspected dip. This stategy works extremely well for Model 3 and Model 5, with investments respectively reaching 6.95 and 6.57 times their initial value. However, liquidating all assets only to buy them back is extremely risky, so one could also employ a strategy of selling only 35% of assets before a predicted decrease and then buying them back at the next predicted increase; this strategy is represented by the purple lines. Even with this less risky strategy, investments under the guidance of Model 3 and Model 5 reach 3.44 and 3.37 times their initial value, which still yields a higher return on investment than a buy-and-hold strategy.

The other four models, however, do not perform as well as Model 3 and Model 5. Even with the risky strategy represented by the blue lines in Figures 23 and 24, the next best models (Model 1 and 2 which, recall, are identical) see the initial investment reach 2.88 times its initial value, which while still better than the buy-and-hold strategy, are nowhere close the gains that Model 3 and Model 5 yield. Furthermore, these increases over the buy-and-hold strategy only happen in the small window from 2020-2021, after the beginning of the dip caused by the COVID-19 pandemic; before this, performance is almost identical to that of the buy-and-hold strategy.

So then, how can Model 3 and Model 5 differ so drastically from the other four models when it comes to return on investment? After all, the largest difference in model accuracy is only 4.19% and only 3.37% when it comes to PPV, so how can such small differences be amplified? Part of the reason lies within the number and timing of predicted increases and decreases, which can be seen in the right panels in Figures 23 and 24. These panels show what happens on the weeks that the model predicted a decrease; on these weeks, investors using the blue line strategy would not be exposed to any increase or decrease in value, and investors using the less risky purple line strategy would only be exposed to 35% of each increase or decrease in value. We see that the magnitude of the mean percent decrease during these weeks is approximately equal to the magnitude of the mean percent increase, so it's not like decreases tend to be stronger or weaker than increases on weeks where a decrease is predicted. However, we do see that Model 3 and Model 5 have a high ratio of decreases avoided to increases avoided; Model 3 avoided 46 decreases and only 24 increases, while Model 5 avoided 45 decreases and only 25 increases. Compare this to the other four models, where the number of decreases avoided is only slightly greater than or approximately equal to the number of increases avoided and we can see that by avoiding significantly more decreases than increases, Model 3 and Model 5 shield investors from decreases while sacrificing only a relatively small number of increases. We also see differences in the timing of these decreases avoided; while Model 3 and Model 5 have the decreases avoided more evenly distributed through time, the other four models tend to have their decreases avoided concentrated in 2020 and 2021. When decreases are avoided earlier in the investment period, the investment increases in value relative to the buy-and-hold strategy, and this investment has a longer time to experience multiplicative increases in value than if decreases are avoided closer to the end of the investment period. Thus, because Model 3 and Model 5 avoid more decreases and do so more consistently throughout the investment period compared to the other four models, the investment value increases faster and more often relative to the other four models or the buy-and-hold strategy.

Discriminant Analysis

Now that we have fit several logistic regression models and found two that performed particularly well, we can examine whether or not using a discriminant analysis approach can yield better-performing models. Because we are working an high-dimensional space, we cannot easily determine whether the decision boundary between the two observed classes (increase and decrease) is linear or nonlinear before fitting any sort of model. Thus, we try both linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA), the former of which uses a linear decision boundary while the latter is more flexible and does not assume equal variance/covariance matrices between the two classes. We can then compare their performance and see which one, if any, is more suitable. The R syntax for the LDA and QDA approaches is very similar to that of using `glm` to fit a logistic model, this time using the functions `lda` and `qda` from the `MASS` package.

```
# Linear discriminant analysis
lda(Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 + BVolume + Prev5GM + Prev5Pct,
    data = SPWeeklyTrain)

# Quadratic discriminant analysis
qda(Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 + BVolume + Prev5GM + Prev5Pct,
    data = SPWeeklyTrain)
```

Just like earlier, there are 255 possible non-interactive models that we can fit, and it is computationally feasible to evaluate every single one of these models. We can build a function `damod.comb`, which is almost identical to `logreg.comb` in terms of functionality, that evaluates each possible model on the training data set and returns a specified statistic such as accuracy, sensitivity, specificity, positive predictive value, or negative predictive value.

```
damod.comb <- function(dataTrain, daType, output, n){
  combn(names(dataTrain)[-c(2:3)], n) %>%
    apply(FUN = paste0, MARGIN = 2, collapse = "+") %>%
    paste0("Direction~", .) %>%
    sapply(formula) %>%
    sapply(FUN = damod, dataTrain = dataTrain, daType = daType, output = output) %>%
    sort() %>%
    return()}
```

We then use this function to select the non-interactive model with the highest training accuracy, first specifying an LDA model for the `daType` argument. This model, which we will call **Model 1**, has an overall training accuracy of approximately 57.62% and correctly predicts 749 out of 1662 movements using a linear combination of the predictors `BVolume`, `Prev5Pct`, `Lag2`, `Lag3`, and `Lag5`. We can also maximise accuracy using this function but by fitting a QDA model rather than an LDA model; doing so yields a model with a slightly higher training accuracy of 58.23%, which we will call **Model 2**. This model only uses the predictors `Prev5GM`, `Lag1`, `Lag3` and `Lag4`.

As we did before with our non-interactive logistic regression models, we can also select LDA/QDA models by maximising PPV rather than accuracy. Doing so gives us an LDA model with a PPV of 57.53% and the exact same structure as Model 1; we will call this **Model 3**. We also obtain a QDA model with a PPV of 59.34%; we will call this **Model 4**; this model has the same predictor variables as Model 2, but with the addition of `BVolume` and `Prev5Pct`.

If we wish to also examine the effectiveness of interactive models, we face the same problem as we did earlier; there are simply too many possible models to evaluate every single one, as we do not have the computational resources to fit and test 68719476735 different models. Thus, rather than identifying a global maximum with certainty, we must perform a heuristic search to maximise accuracy or PPV and hope that we can identify the global maximum, or a local maximum that is approximately equal. Here we use an almost identical version of the forward and backward search algorithms employed on the logistic models, though adapting them to use `lda` and `qda` rather than `glm`.

```r
# Function to remove a term and find new accuracy
damod.lcv1 <- function(i, dataTrain, daType, output){
  predVec[-i] %>%
    paste0(., collapse = "+") %>%
    paste0("Direction~", .) %>%
    formula() %>%
    damod(dataTrain = SPWeeklyTrain, daType = daType, output = output) %>%
    return()}

# Function to return accuracy and new list of predictors after removing a predictor
damod.bwd <- function(predVec, daType, output){
  accs <- sapply(1:length(predVec), FUN = damod.lcv1, dataTrain = SPWeeklyTrain,
                 daType = daType, output = output)
  pv <- predVec[-which.max(accs)]
  return(list(predVec = pv, maxacc = max(accs)))}

# Function to add a term and find new accuracy
damod.lcv2 <- function(i, dataTrain, daType, output){
  c(predVecFwd, predVecRemain[i]) %>%
  paste0(., collapse = "+") %>%
    paste0("Direction~", .) %>%
    formula() %>%
    damod(dataTrain = SPWeeklyTrain, daType = daType, output = output) %>%
    return()}

# Function to return accuracy and new list of predictors after adding a predictor
damod.fwd <- function(predVecRemain, predVecFwd, daType, output){
  accs <- sapply(1:length(predVecRemain), FUN = damod.lcv2, dataTrain = SPWeeklyTrain,
                 daType = daType, output = output)
  pvf <- c(predVecFwd, predVecRemain[which.max(accs)])
  pvr <- predVecRemain[-which.max(accs)]
  return(list(predVecFwd = pvf, predVecRemain = pvr, maxacc = max(accs)))}
```

These functions operate the same as before; the backward-step algorithm starts with a full model and removes terms until accuracy or PPV is maximised, while the forward-step algorith starts with a null model and adds terms until accuracy or PPV is maximised. If we then use these functions to maximise accuracy on an LDA model, we get a 29-term **Model 5** with 59.15% accuracy that was chosen using backward selection and a 8-term **Model 6** with 57.54% accuracy that was chosen using forward selection. We can then apply these functions by maximising accuracy on an QDA model rather than an LDA model, we get a 31-term **Model 7** with 59.46% accuracy that was chosen using backward selection and a 10-term **Model 8** with 58.92% accuracy that was chosen using forward selection.

After using the forward and backward selection algorithms to maximise accuracy, we then use them to maximise PPV. If we use these functions to maximise PPV on an LDA model, we get a 29-term **Model 9** with 59.38% accuracy that was chosen using backward selection and a 8-term **Model 10** with 57.92% accuracy that was chosen using forward selection. We can then apply these functions by maximising PPV on an QDA model rather than an LDA model, we get a 18-term **Model 11** with 49.23% accuracy that was chosen using backward selection and a 13-term **Model 12** with 49.69% accuracy that was chosen using forward selection. Note that unlike the other models, Model 12 and Model 13 have an accuracy of less than 50%, which is worse than simply guessing.

We can then compare all 12 models to see which ones have the highest accuracy and/or PPV, and may thus be the most accurate when it comes to predicting S&P 500 increases or decreases. When it comes to the training data, Model 7 (QDA on backward selection, maximising accuracy) has the highest accuracy at 59.46%, and Model 11 (QDA on backward selection, maximising PPV) has the highest PPV at 75.19%. However, given that we are ultimately trying to make predictions on the test data, we are more interested in the accuracy and PPV there. When our 12 models are applied to the test data, we find that Model 9 (LDA on backward selection, maximising PPV) has the highest accuracy at 63.24%, and Model 7 (QDA on backward selection, maximising accuracy) has the highest PPV at 83.33%. The accuracy and PPV for all 12 models can be found in Figures 25, 26, 27, and 28.

However, when we examine the return on investment by following the predictions of our 12 models, we find that despite a large number of these models having a high accuracy and/or PPV, they still fail to significantly outperform a simple buy-and-hold strategy. This can be seen in Figures 29, 30, 31, and 32. Recall that the buy-and-hold strategy sees the investment increase to 2.30 times its initial value; while all but one of the 12 models outperform the buy-and-hold strategy, only three yield returns that come close to those we observed from the best logistic models: Model 5, Model 9, and Model 7 at 5.59, 5.06, and 4.43 times the initial investment value, respectively. In Model 5 and Model 9, we see the same trend that made some of the logistic models so successful: when selling before a predicted decrease, the number of decreases avoided is greater than the number of increases avoided, and the decreases avoided are spread out over time rather than concentrated near the end of the investment period. For reasons described in the previous section, this combination of characteristics allows for faster growth. Unfortunately, we don't see this with the less successful models, where decreases avoided only slightly outnumber increases avoided and the decreases avoided are concentrated near the end of the investment period.

Close inspection of Figures 30 and 32 will show that the QDA models obtained from backward and forward selection display a very different behaviour from the rest of the models, though. Model 7, Model 8, Model 11, and Model 12 have some of the highest PPVs on the test data, but also tend to predict a rather small number of increases, as is evident by the large number of points on the plots of weeks for which an increase was not predicted. This is also reflected in the graphs of value over time, where the blue curves display horizontal sections that represent long stretches of time where there were no predicted increases, where no trading activity occurs until the next predicted increase. However, Model 7 outperforms Model 8, Model 11, and Model 12 because it completely avoides the large dip caused by the COVID-19 pandemic, which is evident in the centre-right panel of Figure 30 that shows several large decreases avoided during that timeframe. For most of the less-successful models, we also see the model-informed strategies pull ahead of the buy-and-hold strategies around the time of the COVID-19 pandemic, though to a lesser extent than Model 7 does; this same trend was observed with several of our logistic models as well.

Before continuing with our analyses, we note that there are other forms of discriminant analysis that could be used instead of LDA and QDA. For example, mixture discriminant analysis (MDA) relaxes the restriction that each class comes from a Gaussian distribution, and flexible discriminant analysis (FDA) uses non-linear predictor combinations to handle cases of non-normality or non-linearity between variables in a response class[2]. For the sake of brevity, we leave the exercise of fitting these models to the reader.

[2] Examples of how to implement these other types of discriminant analyses in R can be found here.

Support Vector Machines

Decision Trees

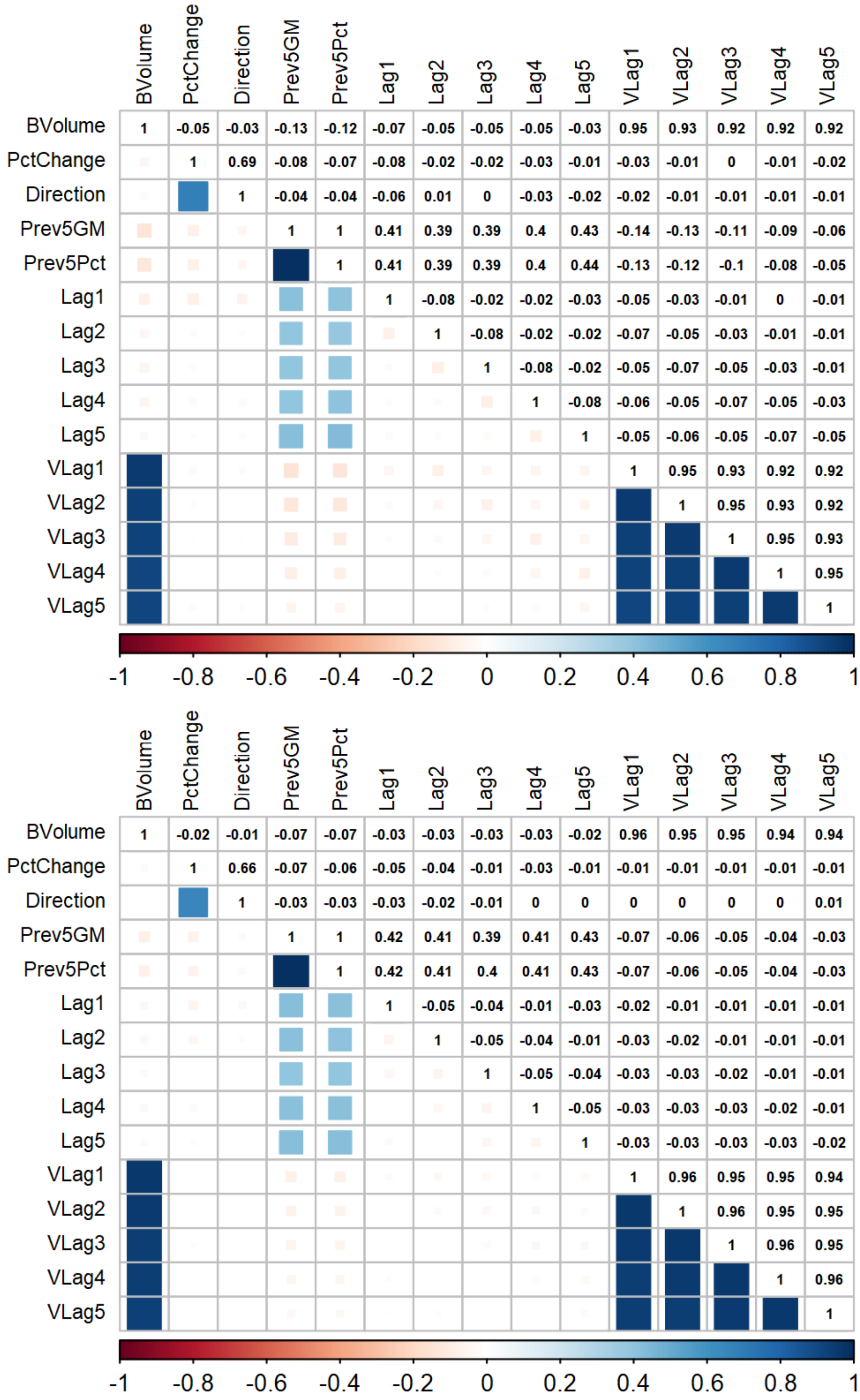
Advanced Methods

Additional Trading Strategies

Conclusions

References

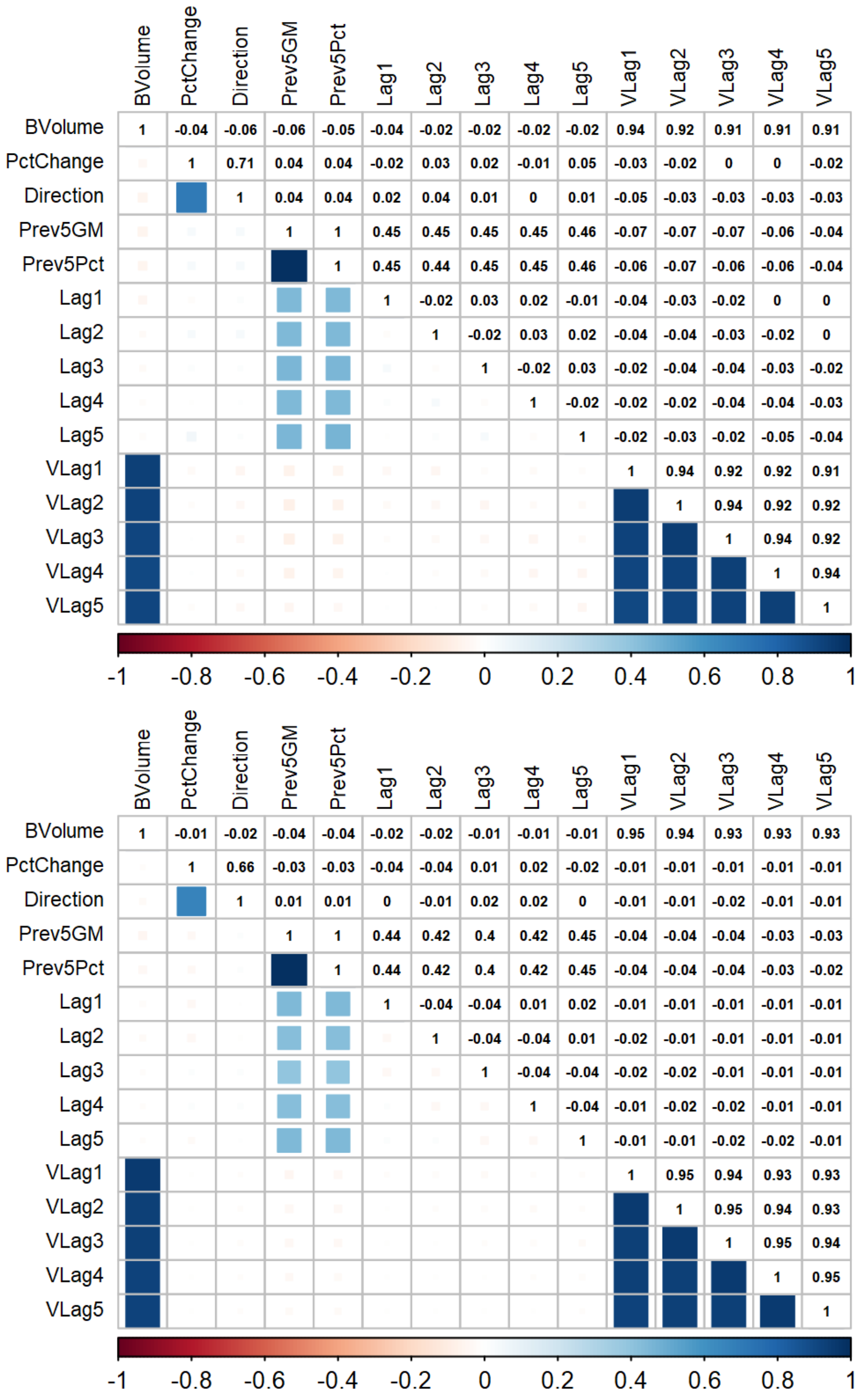
Figures



	BVolume	PctChange	Direction	Prev5GM	Prev5Pct	Lag1	Lag2	Lag3	Lag4	Lag5	VLag1	VLag2	VLag3	VLag4	VLag5
BVolume	1	-0.05	-0.03	-0.13	-0.12	-0.07	-0.05	-0.05	-0.05	-0.03	0.95	0.93	0.92	0.92	0.92
PctChange		1	0.69	-0.08	-0.07	-0.08	-0.02	-0.02	-0.03	-0.01	-0.03	-0.01	0	-0.01	-0.02
Direction			1	-0.04	-0.04	-0.06	0.01	0	-0.03	-0.02	-0.02	-0.01	-0.01	-0.01	-0.01
Prev5GM				1	1	0.41	0.39	0.39	0.4	0.43	-0.14	-0.13	-0.11	-0.09	-0.06
Prev5Pct					1	0.41	0.39	0.39	0.4	0.44	-0.13	-0.12	-0.1	-0.08	-0.05
Lag1						1	-0.08	-0.02	-0.02	-0.03	-0.05	-0.03	-0.01	0	-0.01
Lag2							1	-0.08	-0.02	-0.02	-0.07	-0.05	-0.03	-0.01	-0.01
Lag3								1	-0.08	-0.02	-0.05	-0.07	-0.05	-0.03	-0.01
Lag4									1	-0.08	-0.06	-0.05	-0.07	-0.05	-0.03
Lag5										1	-0.05	-0.06	-0.05	-0.07	-0.05
VLag1											1	0.95	0.93	0.92	0.92
VLag2												1	0.95	0.93	0.92
VLag3													1	0.95	0.93
VLag4														1	0.95
VLag5															1

	BVolume	PctChange	Direction	Prev5GM	Prev5Pct	Lag1	Lag2	Lag3	Lag4	Lag5	VLag1	VLag2	VLag3	VLag4	VLag5
BVolume	1	-0.02	-0.01	-0.07	-0.07	-0.03	-0.03	-0.03	-0.03	-0.02	0.96	0.95	0.95	0.94	0.94
PctChange		1	0.66	-0.07	-0.06	-0.05	-0.04	-0.01	-0.03	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
Direction			1	-0.03	-0.03	-0.03	-0.02	-0.01	0	0	0	0	0	0	0.01
Prev5GM				1	1	0.42	0.41	0.39	0.41	0.43	-0.07	-0.06	-0.05	-0.04	-0.03
Prev5Pct					1	0.42	0.41	0.4	0.41	0.43	-0.07	-0.06	-0.05	-0.04	-0.03
Lag1						1	-0.05	-0.04	-0.01	-0.03	-0.02	-0.01	-0.01	-0.01	-0.01
Lag2							1	-0.05	-0.04	-0.01	-0.03	-0.02	-0.01	-0.01	-0.01
Lag3								1	-0.05	-0.04	-0.03	-0.03	-0.02	-0.01	-0.01
Lag4									1	-0.05	-0.03	-0.03	-0.03	-0.02	-0.01
Lag5										1	-0.03	-0.03	-0.03	-0.03	-0.02
VLag1											1	0.96	0.95	0.95	0.94
VLag2												1	0.96	0.95	0.95
VLag3													1	0.96	0.95
VLag4														1	0.96
VLag5															1

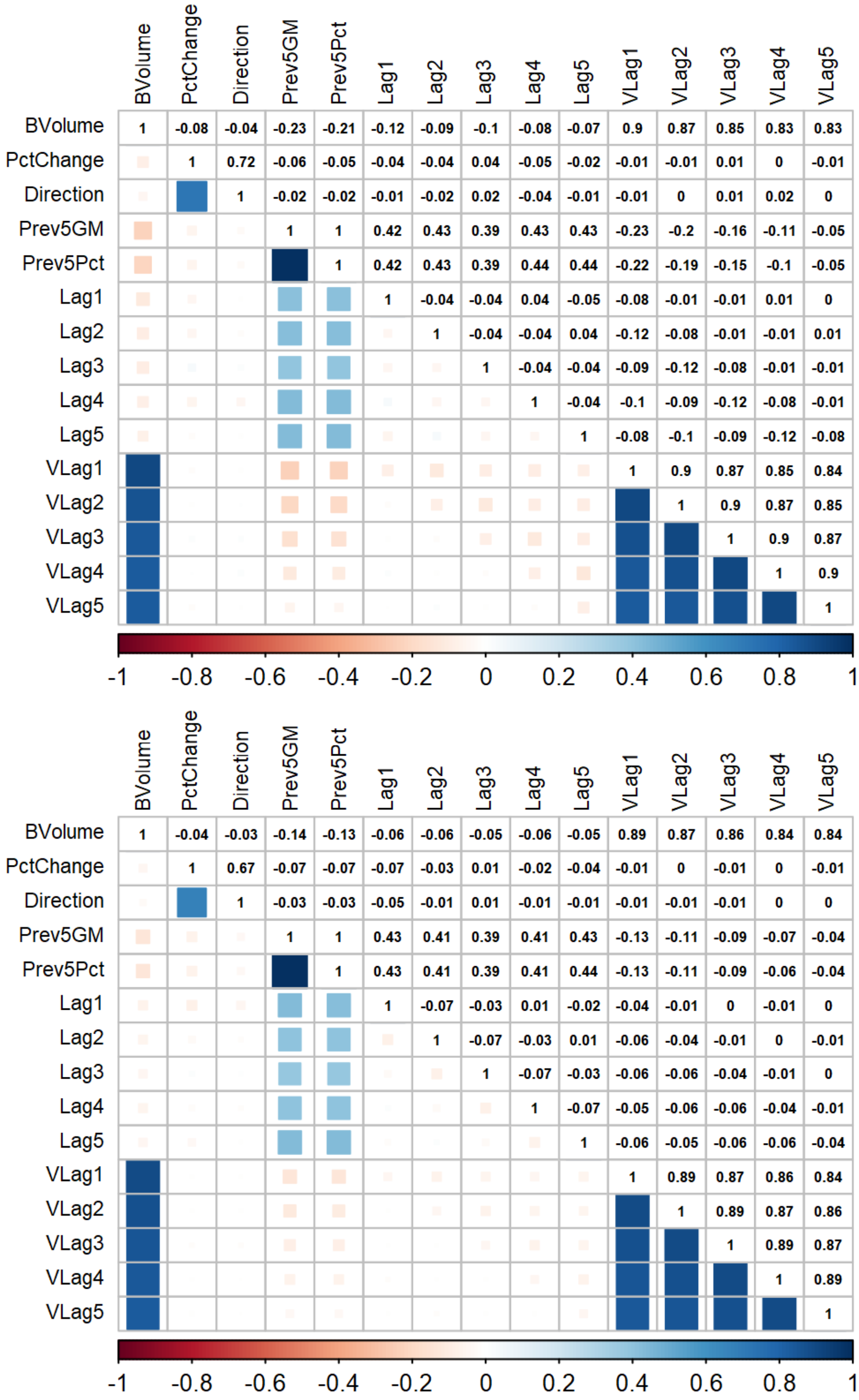
Figure 1: Correlations bewteen volume, percent change, and the various lag variables for the weekly (top) and daily (bottom) data on the S&P 500. The area of squares below the diagonal are proportional to the absolute value of the corresponding correlation coefficients.



	BVolume	PctChange	Direction	Prev5GM	Prev5Pct	Lag1	Lag2	Lag3	Lag4	Lag5	VLag1	VLag2	VLag3	VLag4	VLag5
BVolume	1	-0.04	-0.06	-0.06	-0.05	-0.04	-0.02	-0.02	-0.02	-0.02	0.94	0.92	0.91	0.91	0.91
PctChange		1	0.71	0.04	0.04	-0.02	0.03	0.02	-0.01	0.05	-0.03	-0.02	0	0	-0.02
Direction			1	0.04	0.04	0.02	0.04	0.01	0	0.01	-0.05	-0.03	-0.03	-0.03	-0.03
Prev5GM				1	1	0.45	0.45	0.45	0.45	0.46	-0.07	-0.07	-0.07	-0.06	-0.04
Prev5Pct					1	0.45	0.44	0.45	0.45	0.46	-0.06	-0.07	-0.06	-0.06	-0.04
Lag1						1	-0.02	0.03	0.02	-0.01	-0.04	-0.03	-0.02	0	0
Lag2							1	-0.02	0.03	0.02	-0.04	-0.04	-0.03	-0.02	0
Lag3								1	-0.02	0.03	-0.02	-0.04	-0.04	-0.03	-0.02
Lag4									1	-0.02	-0.02	-0.02	-0.04	-0.04	-0.03
Lag5										1	-0.02	-0.03	-0.02	-0.05	-0.04
VLag1											1	0.94	0.92	0.92	0.91
VLag2												1	0.94	0.92	0.92
VLag3													1	0.94	0.92
VLag4														1	0.94
VLag5															1

	BVolume	PctChange	Direction	Prev5GM	Prev5Pct	Lag1	Lag2	Lag3	Lag4	Lag5	VLag1	VLag2	VLag3	VLag4	VLag5
BVolume	1	-0.01	-0.02	-0.04	-0.04	-0.02	-0.02	-0.01	-0.01	-0.01	0.95	0.94	0.93	0.93	0.93
PctChange		1	0.66	-0.03	-0.03	-0.04	-0.04	0.01	0.02	-0.02	-0.01	-0.01	-0.01	-0.01	-0.01
Direction			1	0.01	0.01	0	-0.01	0.02	0.02	0	-0.01	-0.01	-0.02	-0.01	-0.01
Prev5GM				1	1	0.44	0.42	0.4	0.42	0.45	-0.04	-0.04	-0.04	-0.03	-0.03
Prev5Pct					1	0.44	0.42	0.4	0.42	0.45	-0.04	-0.04	-0.04	-0.03	-0.02
Lag1						1	-0.04	-0.04	0.01	0.02	-0.01	-0.01	-0.01	-0.01	-0.01
Lag2							1	-0.04	-0.04	0.01	-0.02	-0.01	-0.01	-0.01	-0.01
Lag3								1	-0.04	-0.04	-0.02	-0.02	-0.01	-0.01	-0.01
Lag4									1	-0.04	-0.01	-0.02	-0.02	-0.01	-0.01
Lag5										1	-0.01	-0.01	-0.02	-0.02	-0.01
VLag1											1	0.95	0.94	0.93	0.93
VLag2												1	0.95	0.94	0.93
VLag3													1	0.95	0.94
VLag4														1	0.95
VLag5															1

Figure 2: Correlations bewteen volume, percent change, and the various lag variables for the weekly (top) and daily (bottom) data on the NASDAQ Composite. The area of squares below the diagonal are proportional to the absolute value of the corresponding correlation coefficients.



	BVolume	PctChange	Direction	Prev5GM	Prev5Pct	Lag1	Lag2	Lag3	Lag4	Lag5	VLag1	VLag2	VLag3	VLag4	VLag5
BVolume	1	-0.08	-0.04	-0.23	-0.21	-0.12	-0.09	-0.1	-0.08	-0.07	0.9	0.87	0.85	0.83	0.83
PctChange		1	0.72	-0.06	-0.05	-0.04	-0.04	0.04	-0.05	-0.02	-0.01	-0.01	0.01	0	-0.01
Direction			1	-0.02	-0.02	-0.01	-0.02	0.02	-0.04	-0.01	-0.01	0	0.01	0.02	0
Prev5GM				1	1	0.42	0.43	0.39	0.43	0.43	-0.23	-0.2	-0.16	-0.11	-0.05
Prev5Pct					1	0.42	0.43	0.39	0.44	0.44	-0.22	-0.19	-0.15	-0.1	-0.05
Lag1						1	-0.04	-0.04	0.04	-0.05	-0.08	-0.01	-0.01	0.01	0
Lag2							1	-0.04	-0.04	0.04	-0.12	-0.08	-0.01	-0.01	0.01
Lag3								1	-0.04	-0.04	-0.09	-0.12	-0.08	-0.01	-0.01
Lag4									1	-0.04	-0.1	-0.09	-0.12	-0.08	-0.01
Lag5										1	-0.08	-0.1	-0.09	-0.12	-0.08
VLag1											1	0.9	0.87	0.85	0.84
VLag2												1	0.9	0.87	0.85
VLag3													1	0.9	0.87
VLag4														1	0.9
VLag5															1

	BVolume	PctChange	Direction	Prev5GM	Prev5Pct	Lag1	Lag2	Lag3	Lag4	Lag5	VLag1	VLag2	VLag3	VLag4	VLag5
BVolume	1	-0.04	-0.03	-0.14	-0.13	-0.06	-0.06	-0.05	-0.06	-0.05	0.89	0.87	0.86	0.84	0.84
PctChange		1	0.67	-0.07	-0.07	-0.07	-0.03	0.01	-0.02	-0.04	-0.01	0	-0.01	0	-0.01
Direction			1	-0.03	-0.03	-0.05	-0.01	0.01	-0.01	-0.01	-0.01	-0.01	-0.01	0	0
Prev5GM				1	1	0.43	0.41	0.39	0.41	0.43	-0.13	-0.11	-0.09	-0.07	-0.04
Prev5Pct					1	0.43	0.41	0.39	0.41	0.44	-0.13	-0.11	-0.09	-0.06	-0.04
Lag1						1	-0.07	-0.03	0.01	-0.02	-0.04	-0.01	0	-0.01	0
Lag2							1	-0.07	-0.03	0.01	-0.06	-0.04	-0.01	0	-0.01
Lag3								1	-0.07	-0.03	-0.06	-0.06	-0.04	-0.01	0
Lag4									1	-0.07	-0.05	-0.06	-0.06	-0.04	-0.01
Lag5										1	-0.06	-0.05	-0.06	-0.06	-0.04
VLag1											1	0.89	0.87	0.86	0.84
VLag2												1	0.89	0.87	0.86
VLag3													1	0.89	0.87
VLag4														1	0.89
VLag5															1

Figure 3: Correlations bewteen volume, percent change, and the various lag variables for the weekly (top) and daily (bottom) data on the Dow Jones Industrial Average. The area of squares below the diagonal are proportional to the absolute value of the corresponding correlation coefficients.

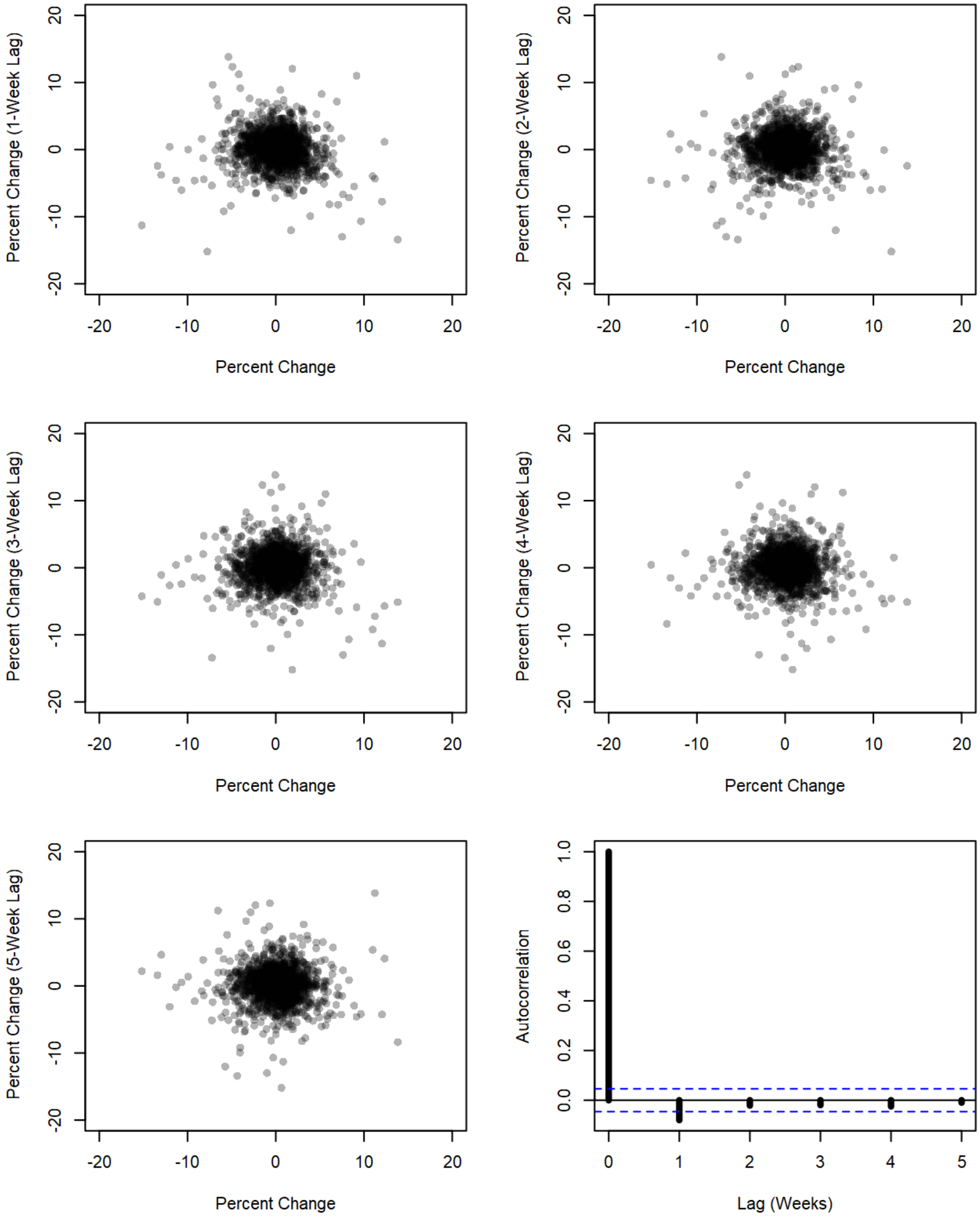


Figure 4: Plots of S&P 500 weekly percent change against percent change of up to 5 weeks prior. Autocorrelation in daily percent change corresponding to 1-5 day lag times is shown in the bottom-right panel; dashed blue lines represent the 95% confidence interval for the autocorrelation.

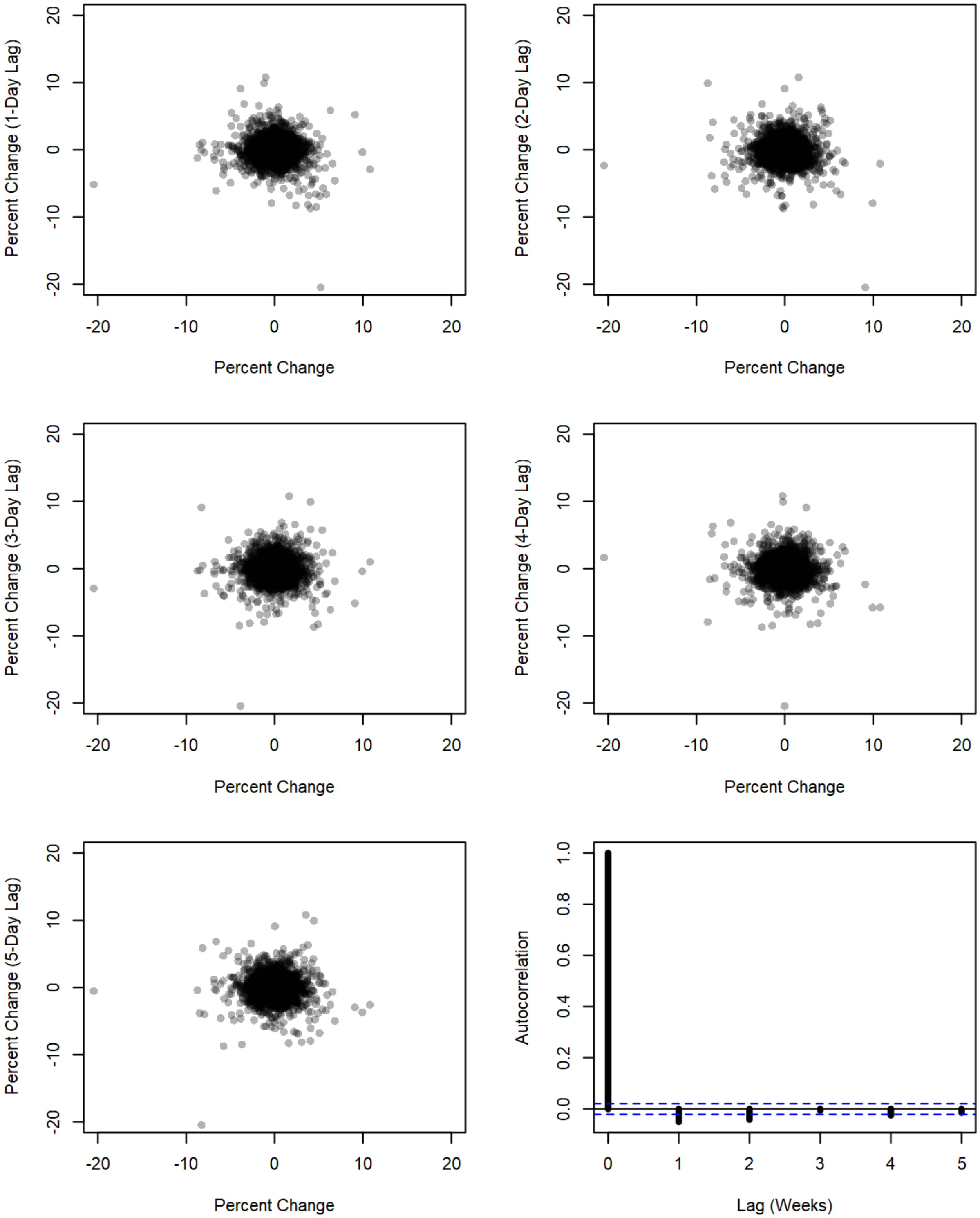


Figure 5: Plots of S&P 500 daily percent change against percent change of up to 5 days prior. Autocorrelation in daily percent change corresponding to 1-5 day lag times is shown in the bottom-right panel; dashed blue lines represent the 95% confidence interval for the autocorrelation.

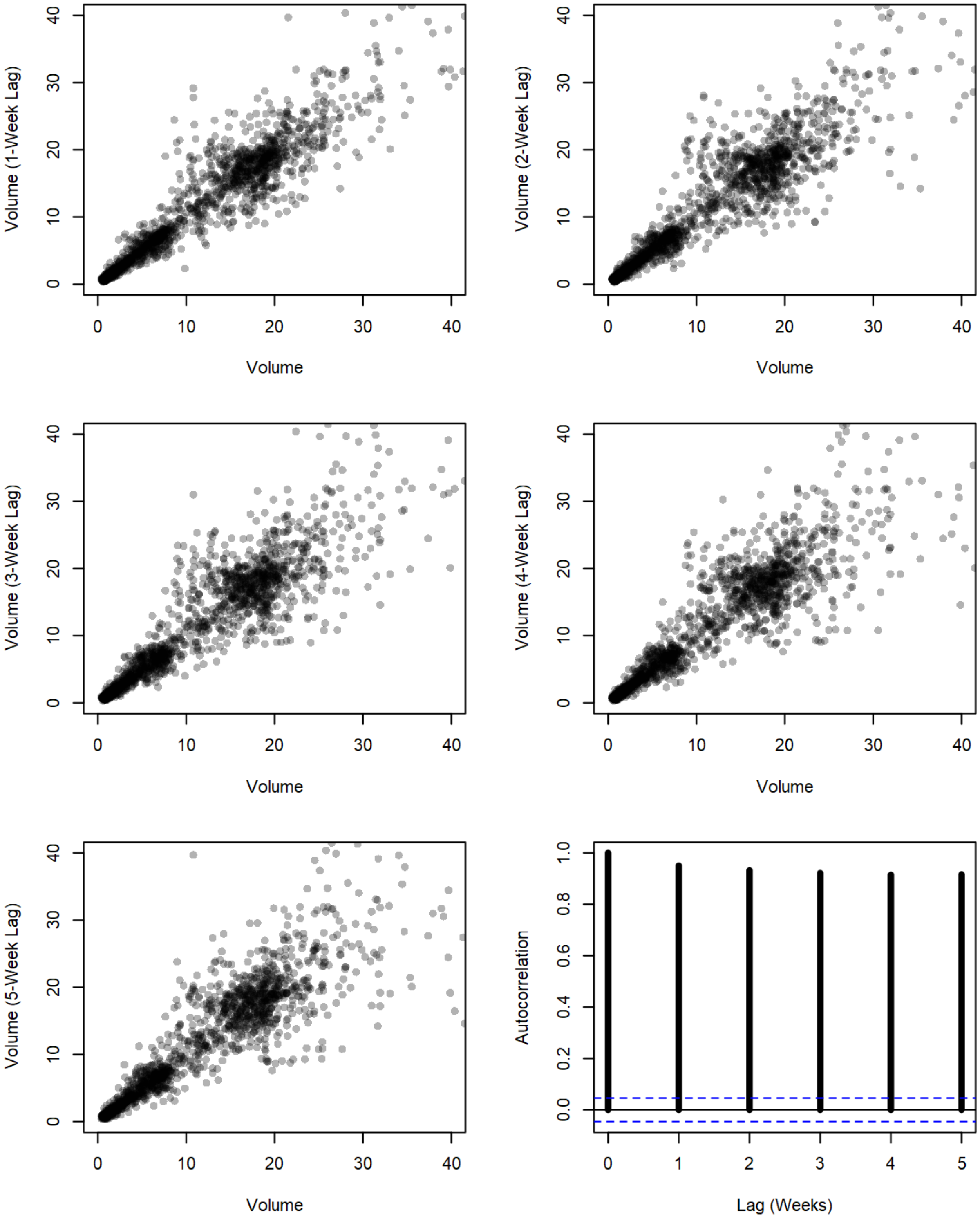


Figure 6: Plots of S&P 500 weekly volume against volume of up to 5 weeks prior. Autocorrelation in volume corresponding to 1-5 week lag times is shown in the bottom-right panel; dashed blue lines represent the 95% confidence interval for the autocorrelation.

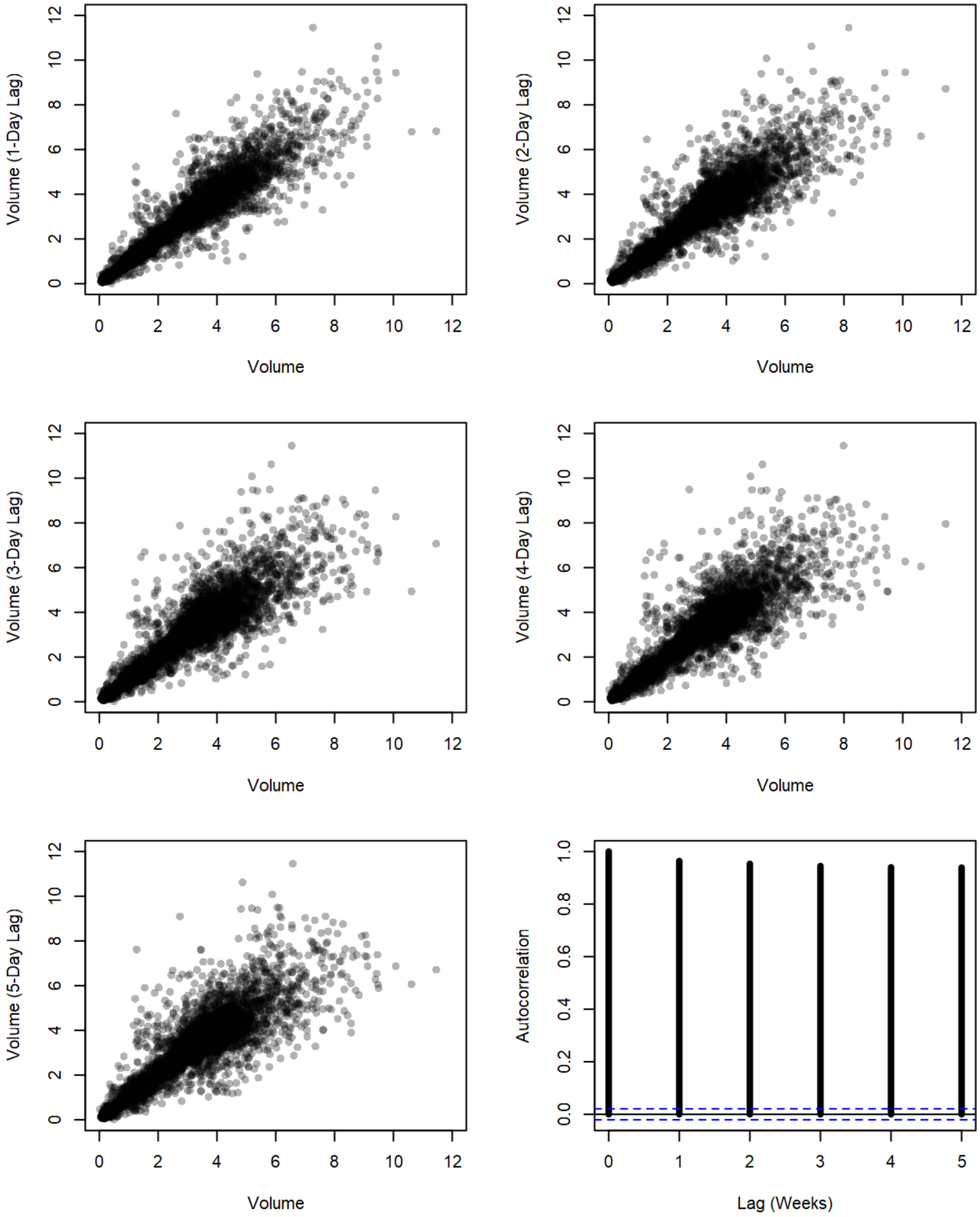


Figure 7: Plots of S&P 500 daily volume against volume of up to 5 days prior. Autocorrelation in volume corresponding to 1-5 day lag times is shown in the bottom-right panel; dashed blue lines represent the 95% confidence interval for the autocorrelation.

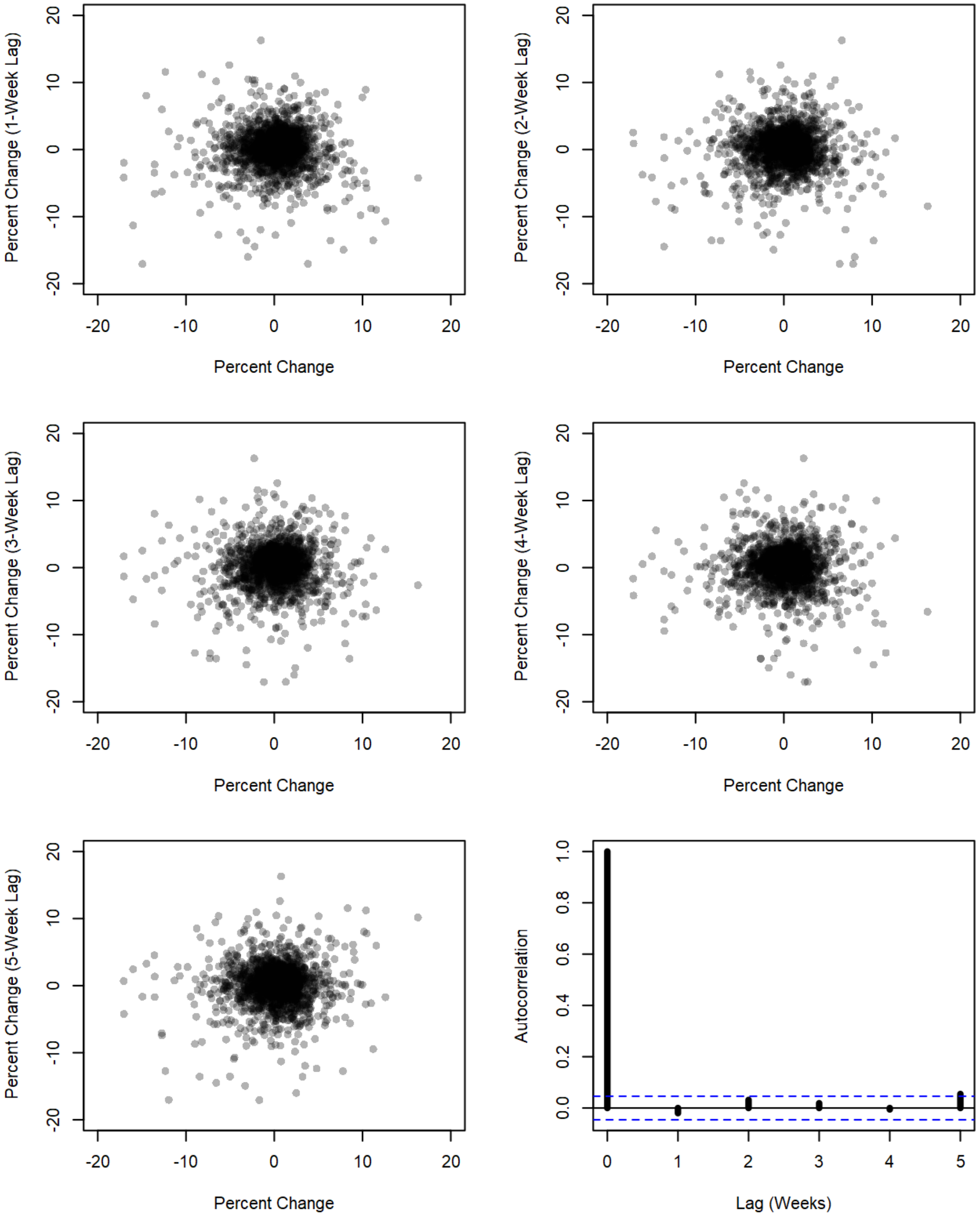


Figure 8: Plots of NASDAQ Composite weekly percent change against percent change of up to 5 weeks prior. Autocorrelation in weekly percent change corresponding to 1-5 week lag times is shown in the bottom-right panel; dashed blue lines represent the 95% confidence interval for the autocorrelation.

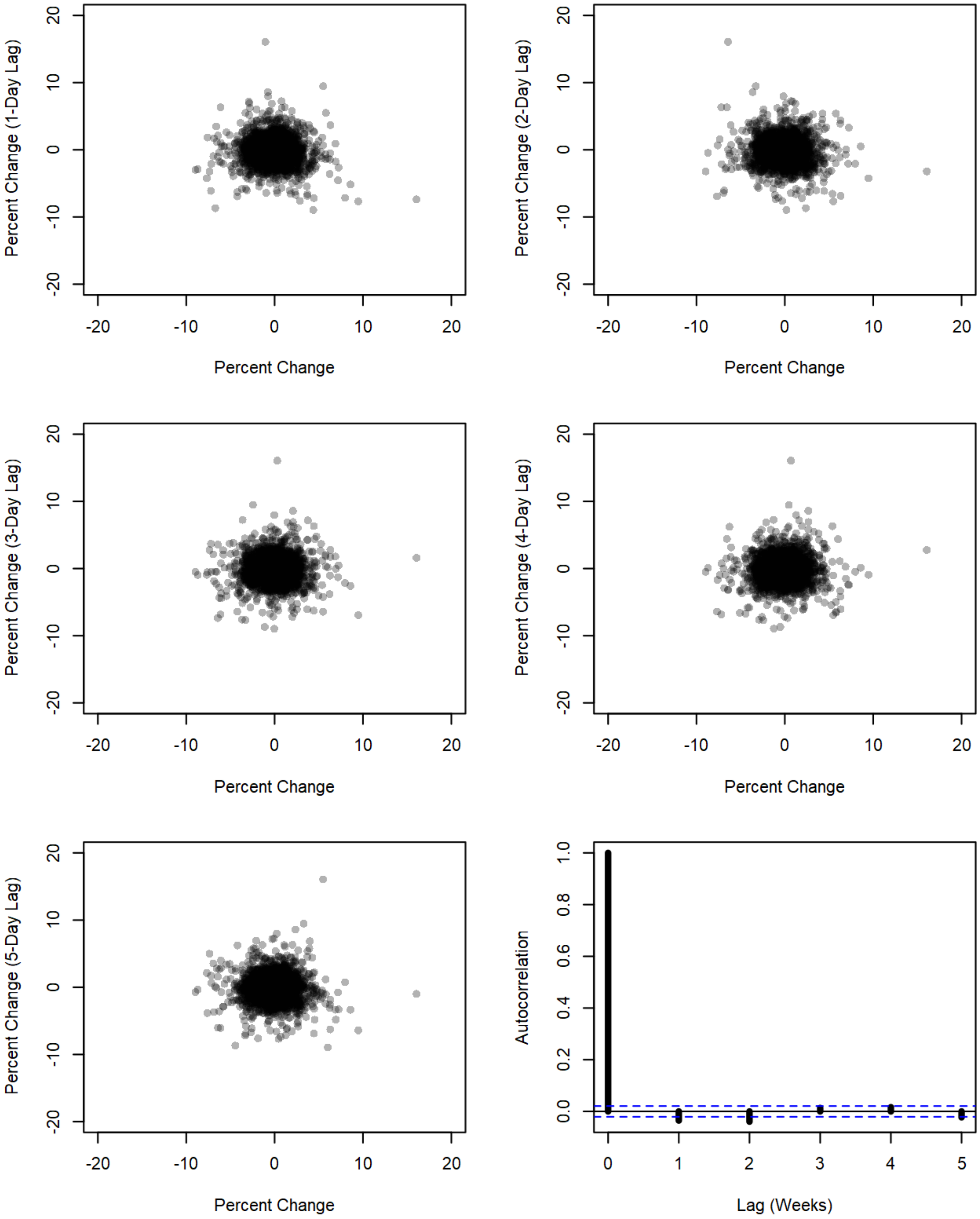


Figure 9: Plots of NASDAQ Composite daily percent change against percent change of up to 5 days prior. Autocorrelation in daily percent change corresponding to 1-5 day lag times is shown in the bottom-right panel; dashed blue lines represent the 95% confidence interval for the autocorrelation.

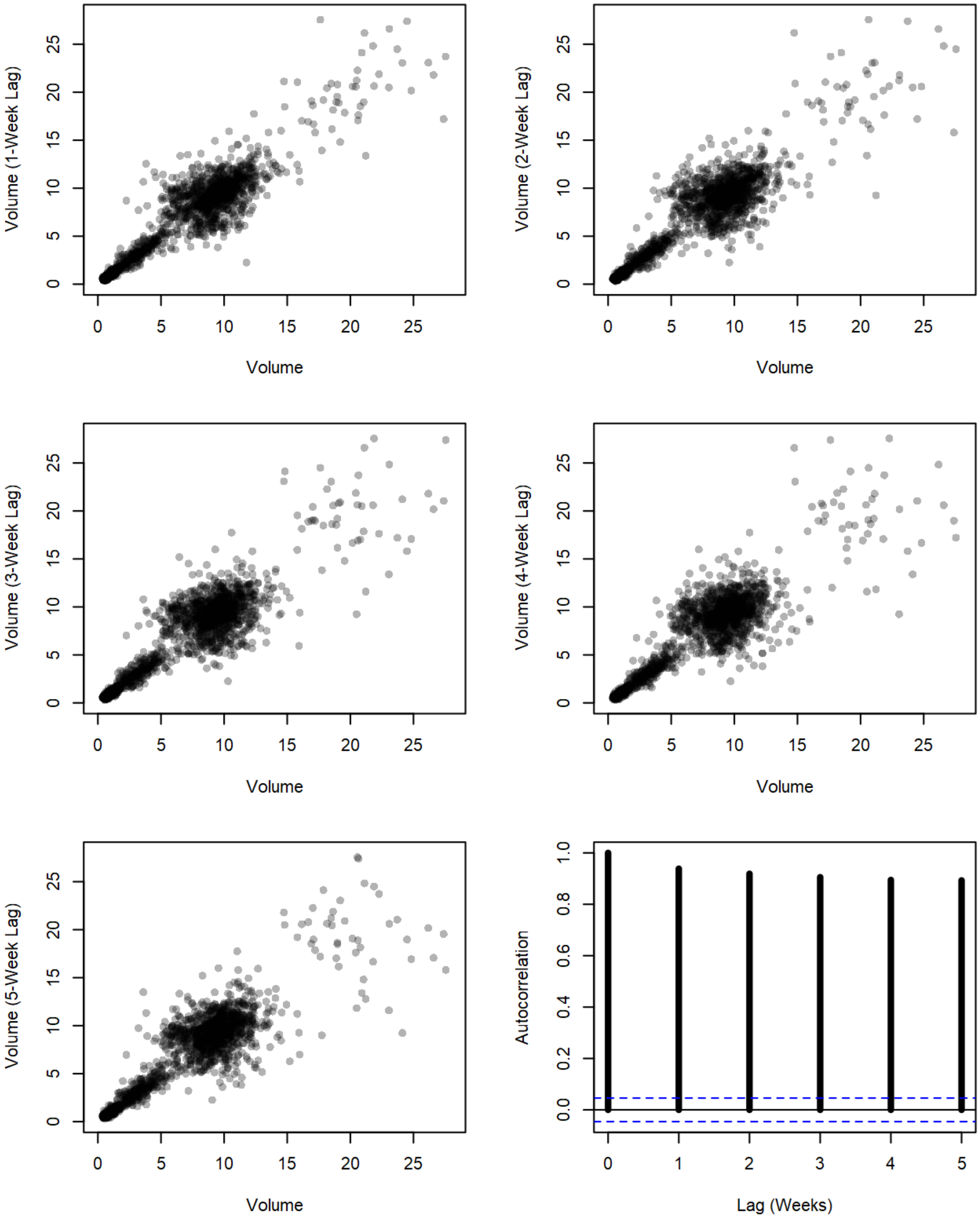


Figure 10: Plots of NASDAQ Composite weekly volume against volume of up to 5 weeks prior. Autocorrelation in volume corresponding to 1-5 week lag times is shown in the bottom-right panel; dashed blue lines represent the 95% confidence interval for the autocorrelation.

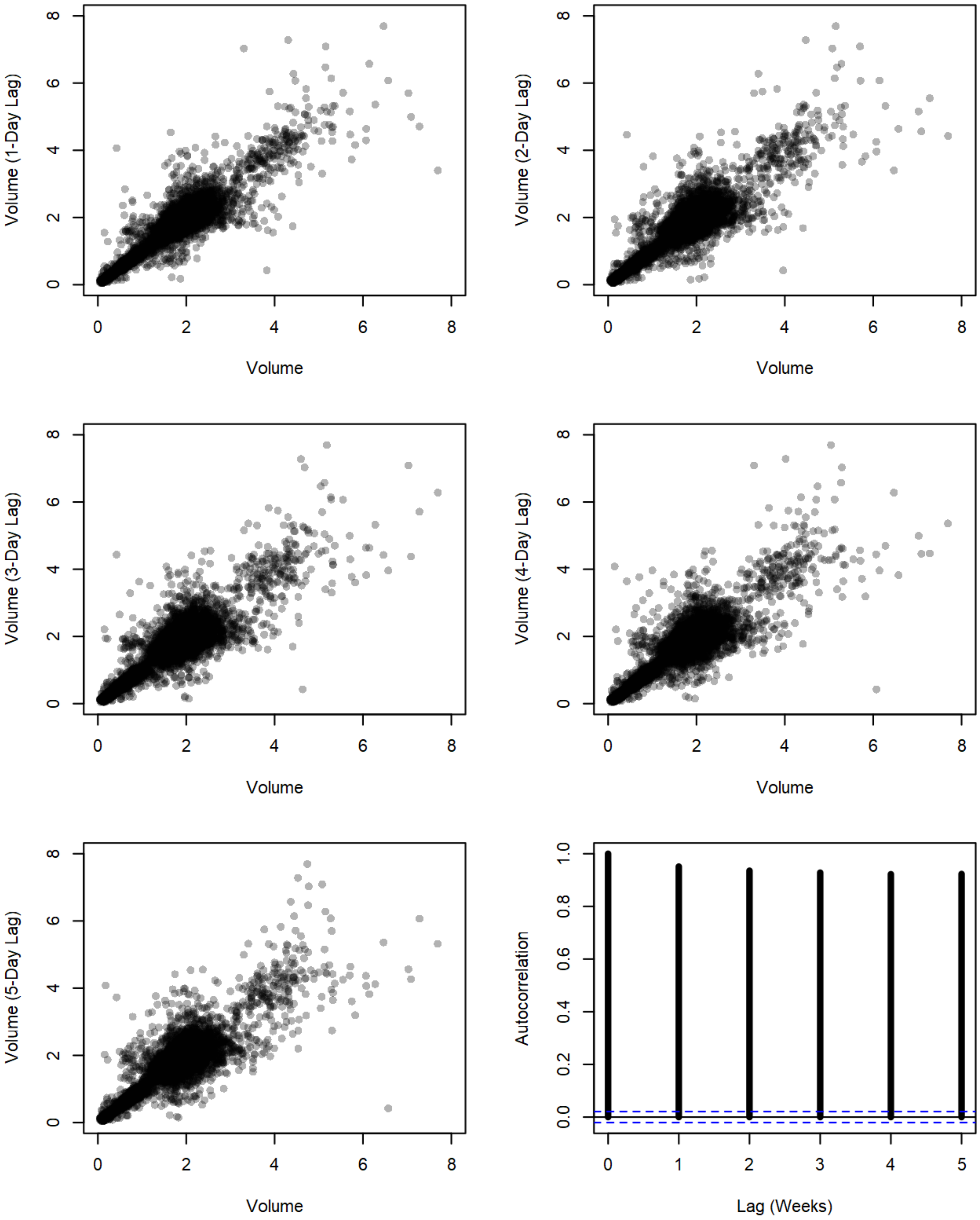


Figure 11: Plots of NASDAQ Composite daily volume against volume of up to 5 days prior. Autocorrelation in volume corresponding to 1-5 day lag times is shown in the bottom-right panel; dashed blue lines represent the 95% confidence interval for the autocorrelation.

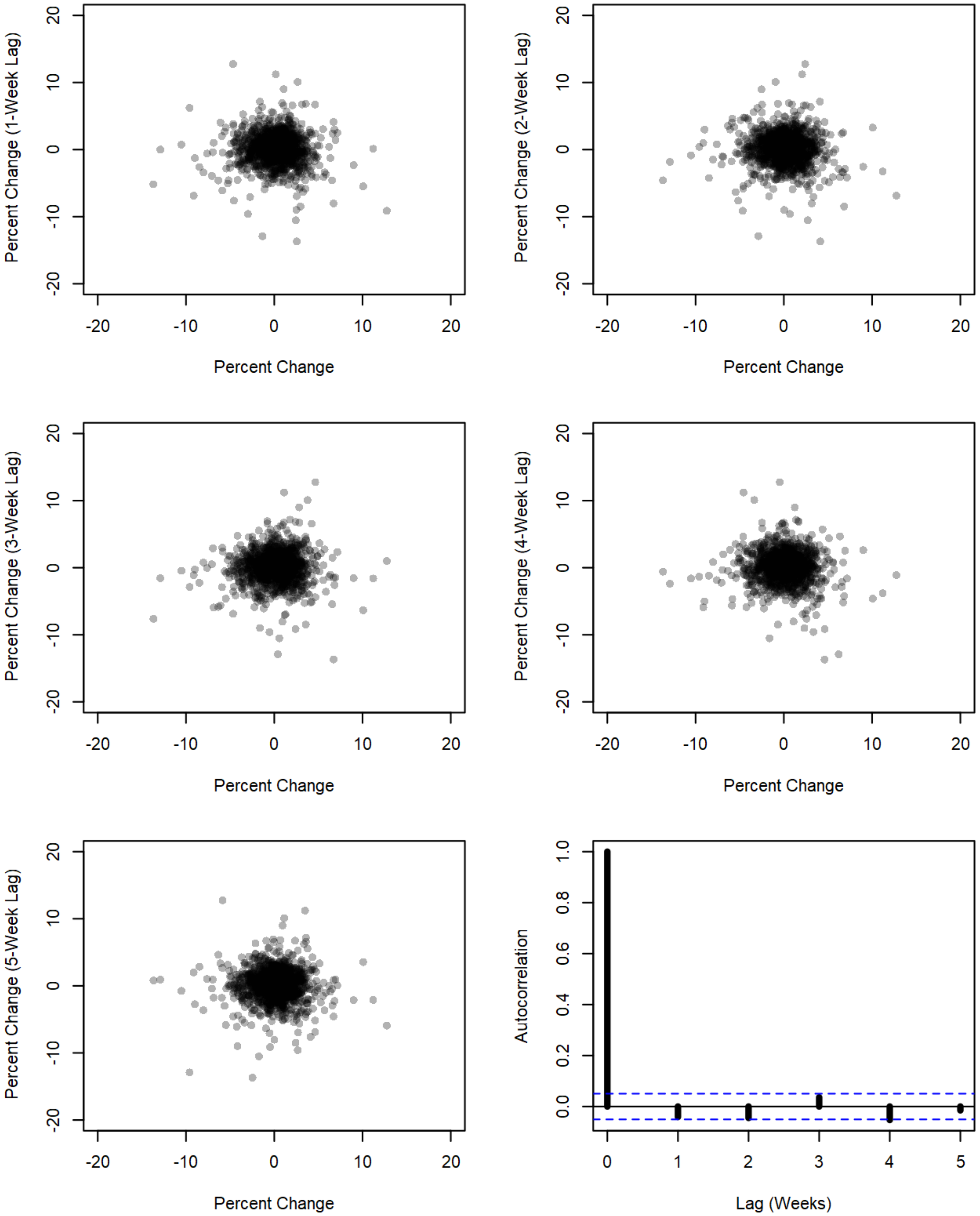


Figure 12: Plots of Dow Jones Industrial Average weekly percent change against percent change of up to 5 weeks prior. Autocorrelation in weekly percent change corresponding to 1-5 week lag times is shown in the bottom-right panel; dashed blue lines represent the 95% confidence interval for the autocorrelation.

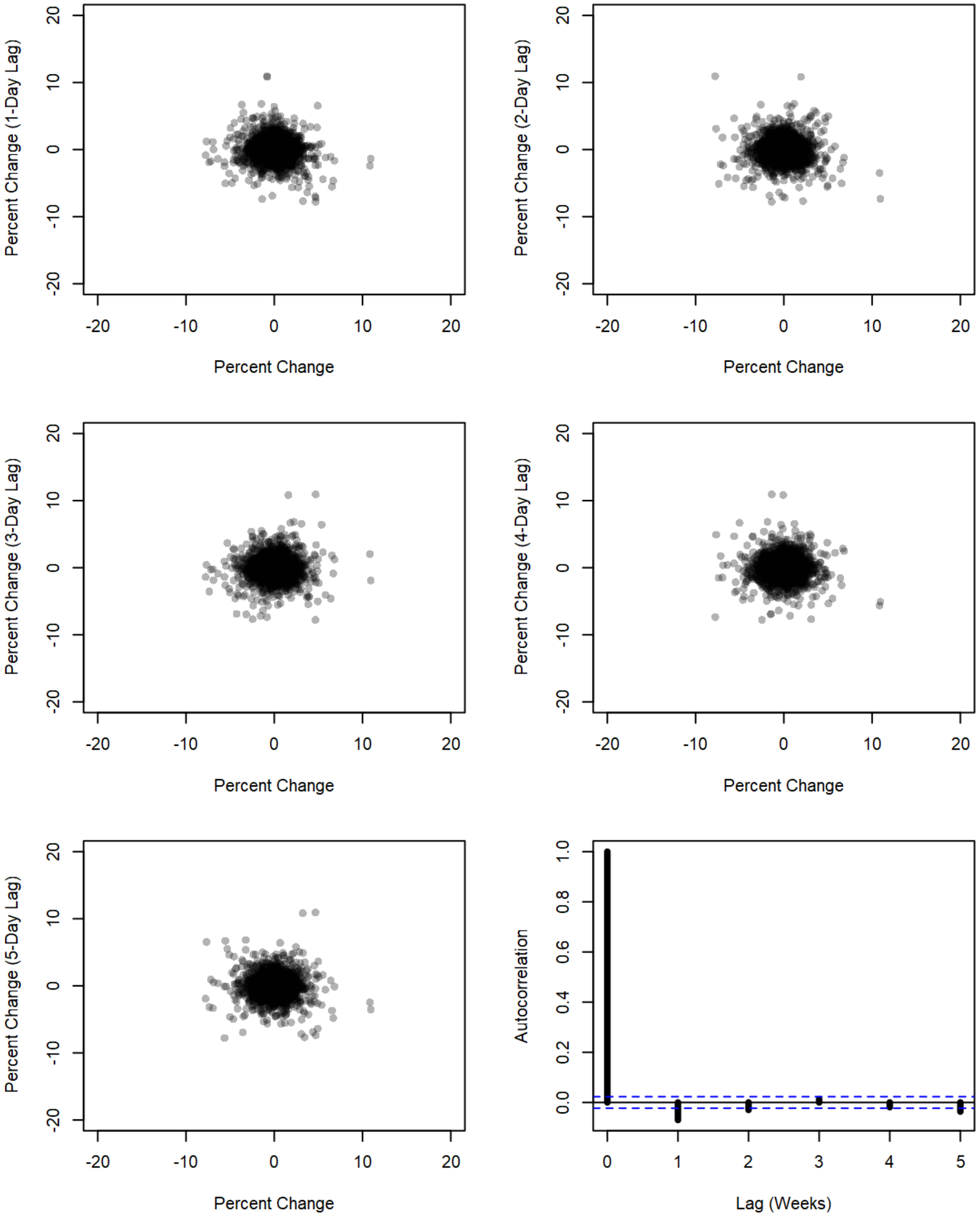


Figure 13: Plots of Dow Jones Industrial Average daily percent change against percent change of up to 5 days prior. Autocorrelation in daily percent change corresponding to 1-5 day lag times is shown in the bottom-right panel; dashed blue lines represent the 95% confidence interval for the autocorrelation.

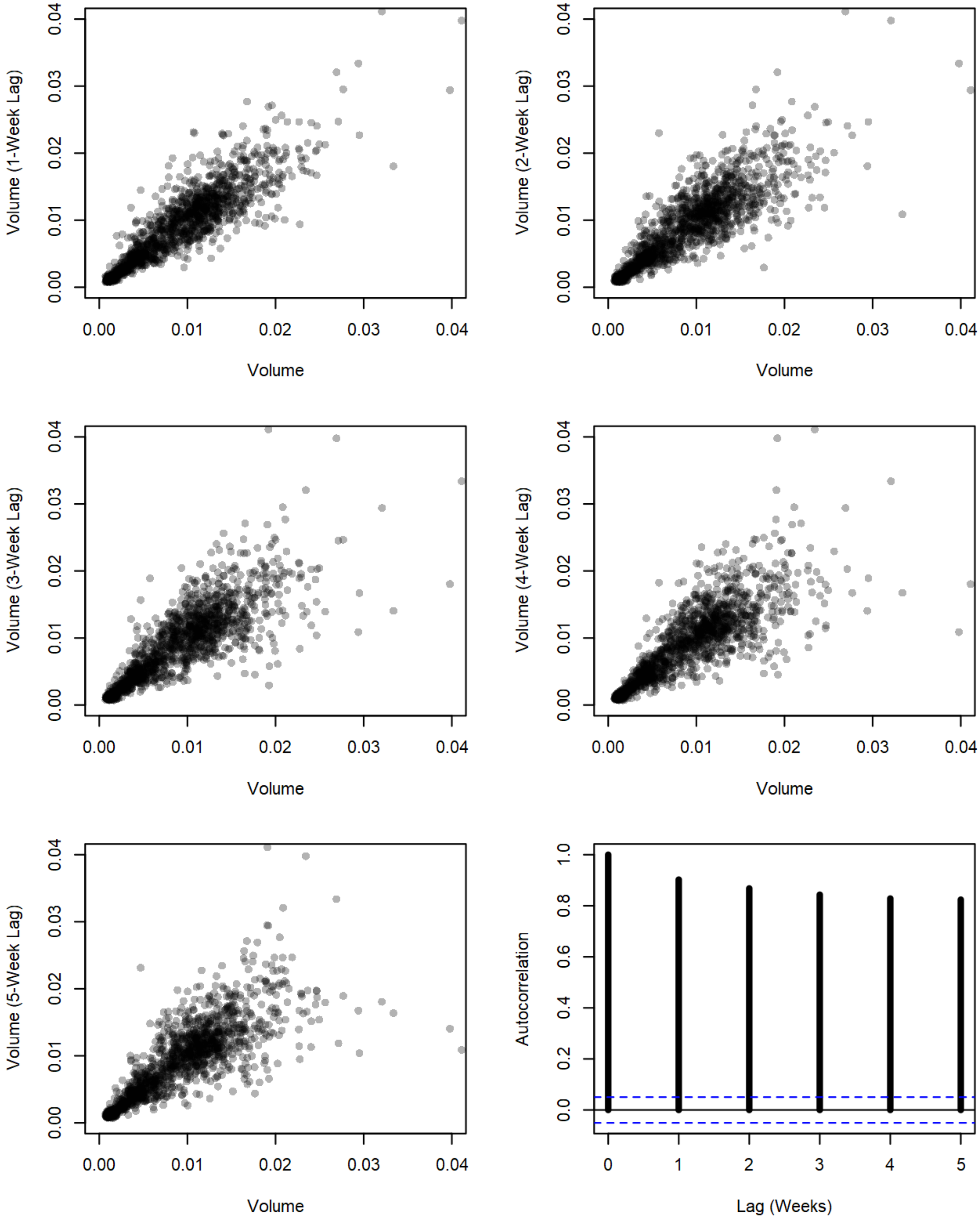


Figure 14: Plots of Dow Jones Industrial Average weekly volume against volume of up to 5 weeks prior. Autocorrelation in volume corresponding to 1-5 week lag times is shown in the bottom-right panel; dashed blue lines represent the 95% confidence interval for the autocorrelation.

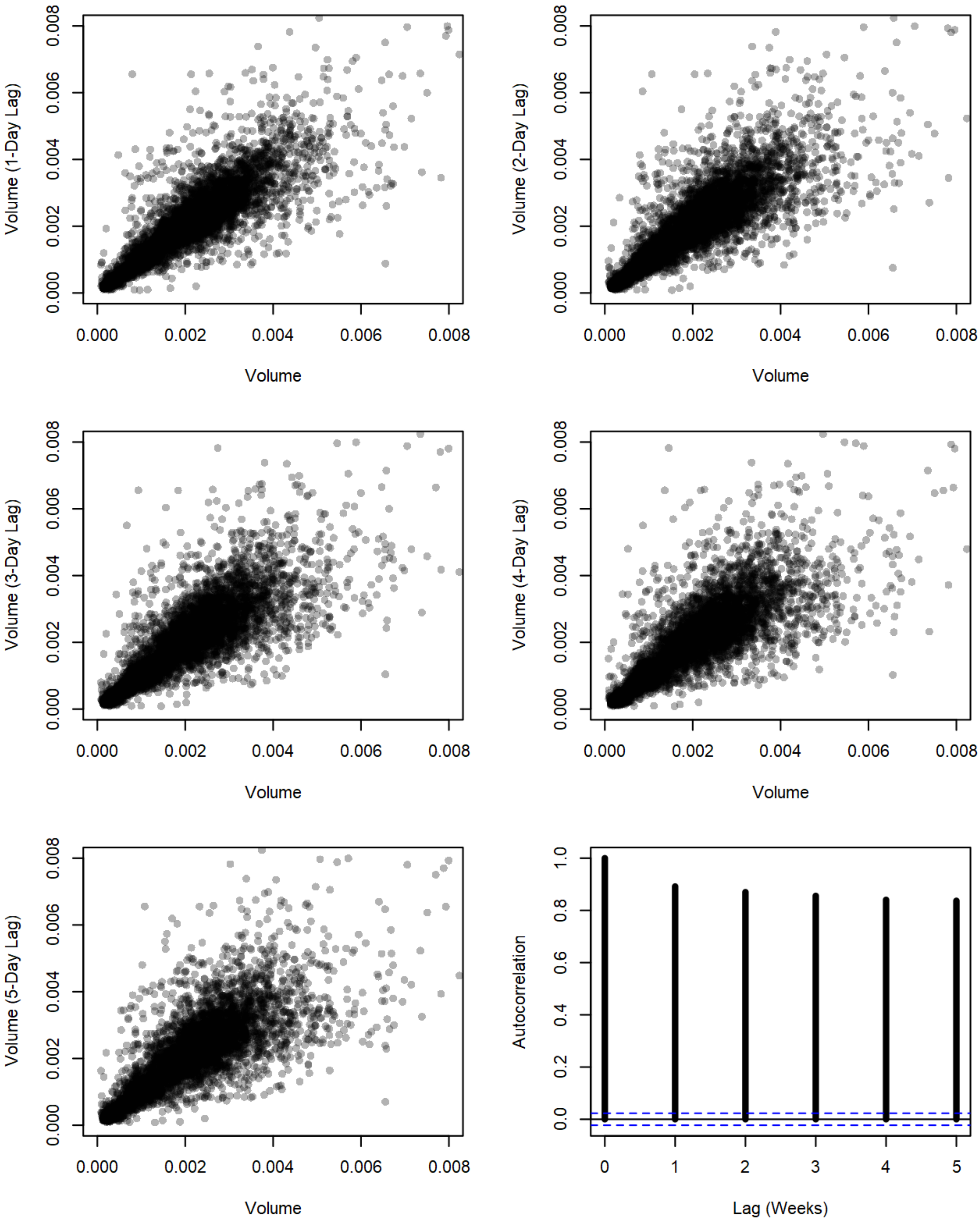


Figure 15: Plots of Dow Jones Industrial Average daily volume against volume of up to 5 days prior. Autocorrelation in volume corresponding to 1-5 day lag times is shown in the bottom-right panel; dashed blue lines represent the 95% confidence interval for the autocorrelation.

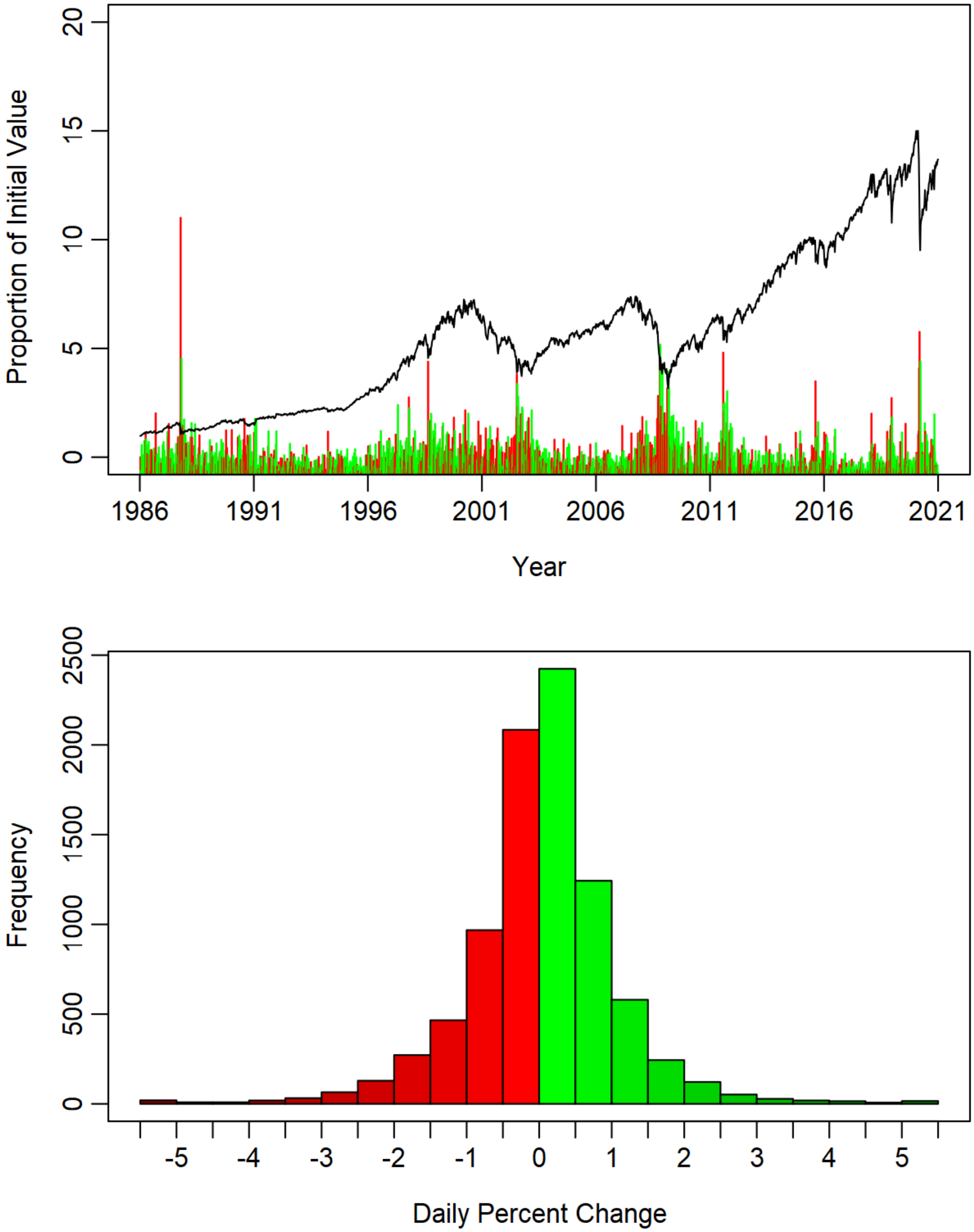


Figure 16: Relative value of the S&P 500 over time compared to its 1986 value (top) and distribution of daily percent changes (bottom). Note that in the bottom panel, the leftmost bin represents all days with a percent change less than -5, and rightmost bin represents all days with a percent change greater than 5. Note that the top chart does not account for after-hours gains or losses.

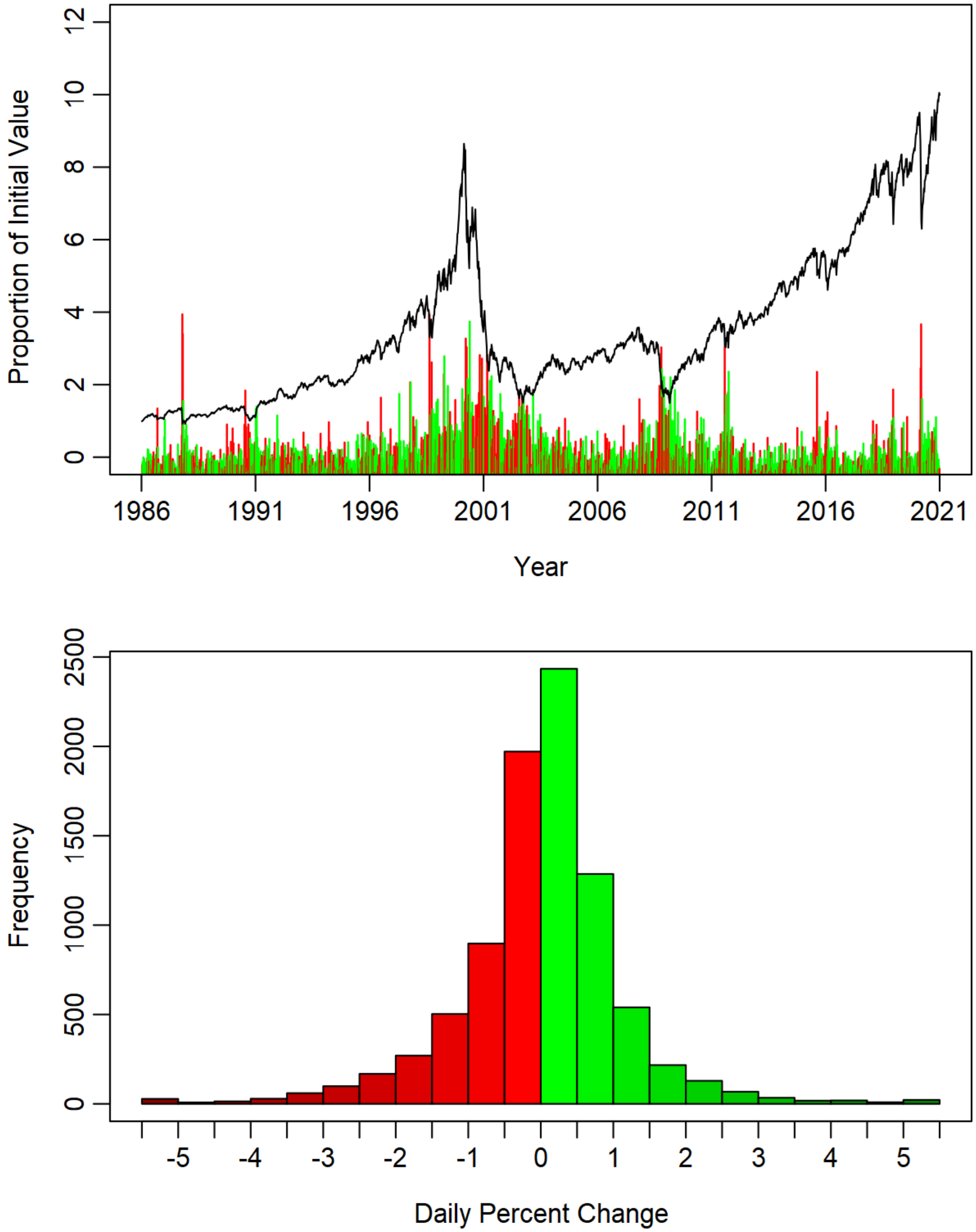


Figure 17: Relative value of the NASDAQ Composite over time compared to its 1986 value (top) and distribution of daily percent changes (bottom). Note that in the bottom panel, the leftmost bin represents all days with a percent change less than -5, and rightmost bin represents all days with a percent change greater than 5. Note that the top chart does not account for after-hours gains or losses.

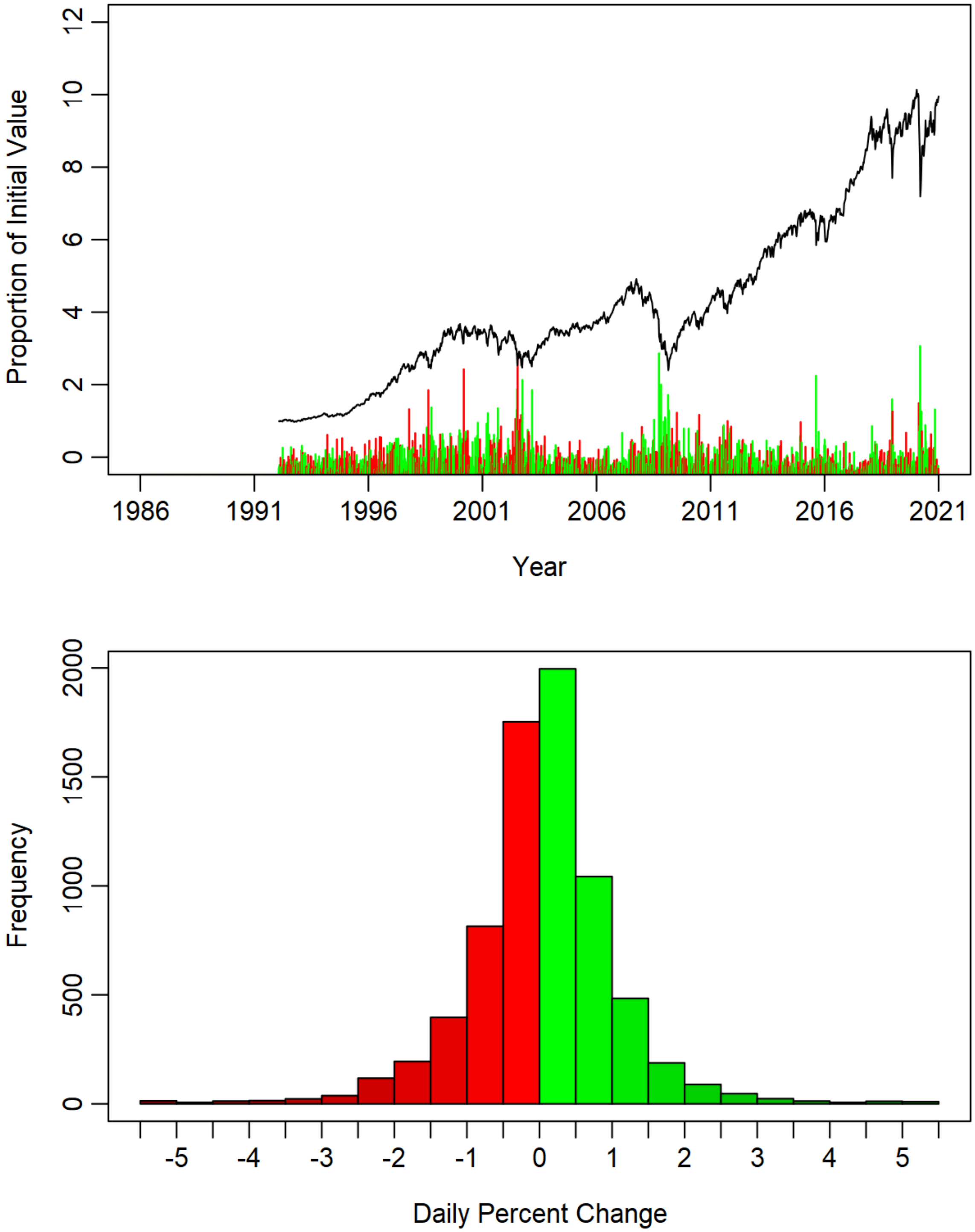


Figure 18: Relative value of the Dow Jones Industrial Average over time compared to its 1992 value (top) and distribution of daily percent changes (bottom). Note that in the bottom panel, the leftmost bin represents all days with a percent change less than -5, and rightmost bin represents all days with a percent change greater than 5. Note that the top chart does not account for after-hours gains or losses.

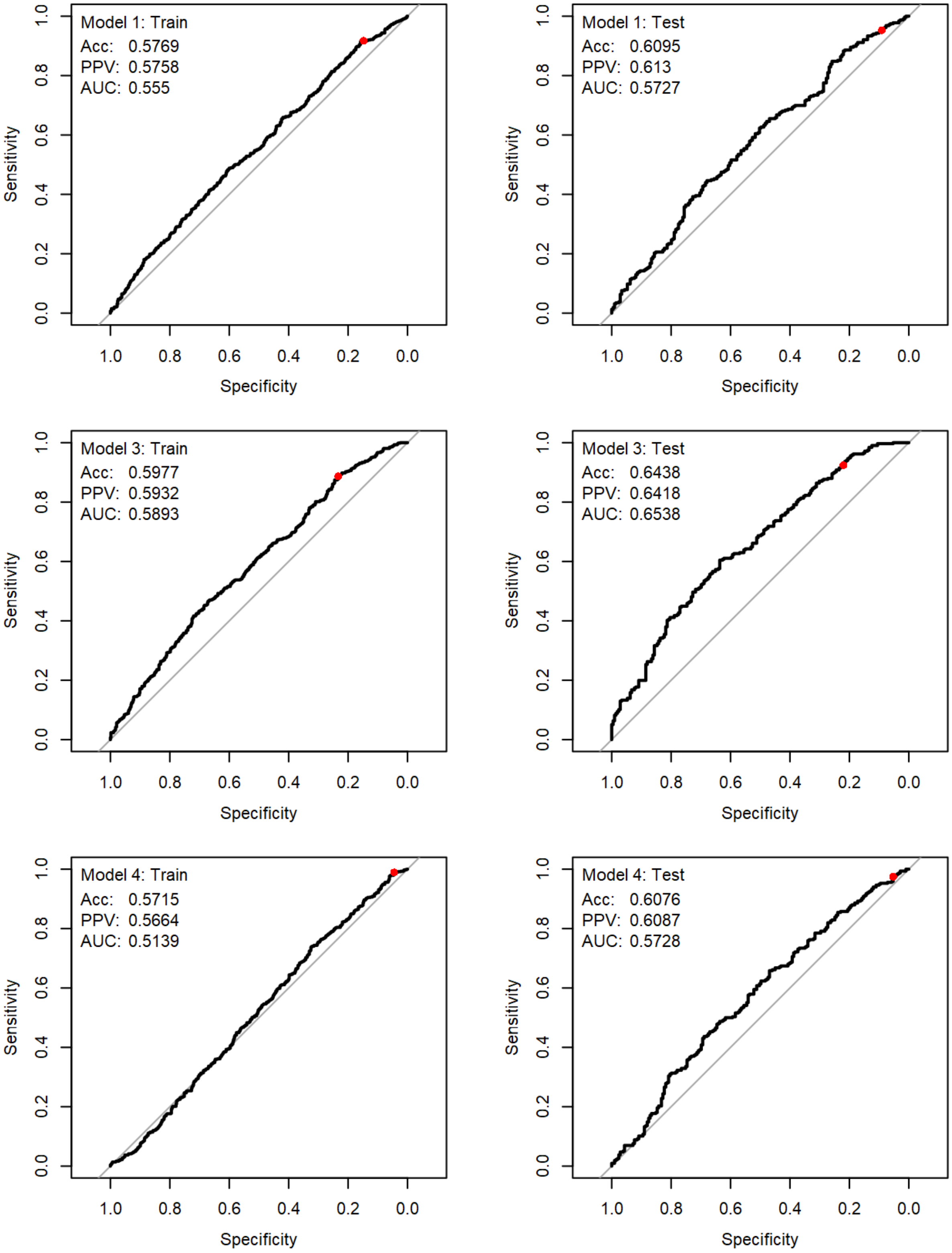


Figure 19: ROC curves for the logistic regression models selected by maximising accuracy. Accuracy, positive predictive value (PPV), and area under the curve (AUC) are listed for each plot; the red dot represents model sensitivity and specificity.

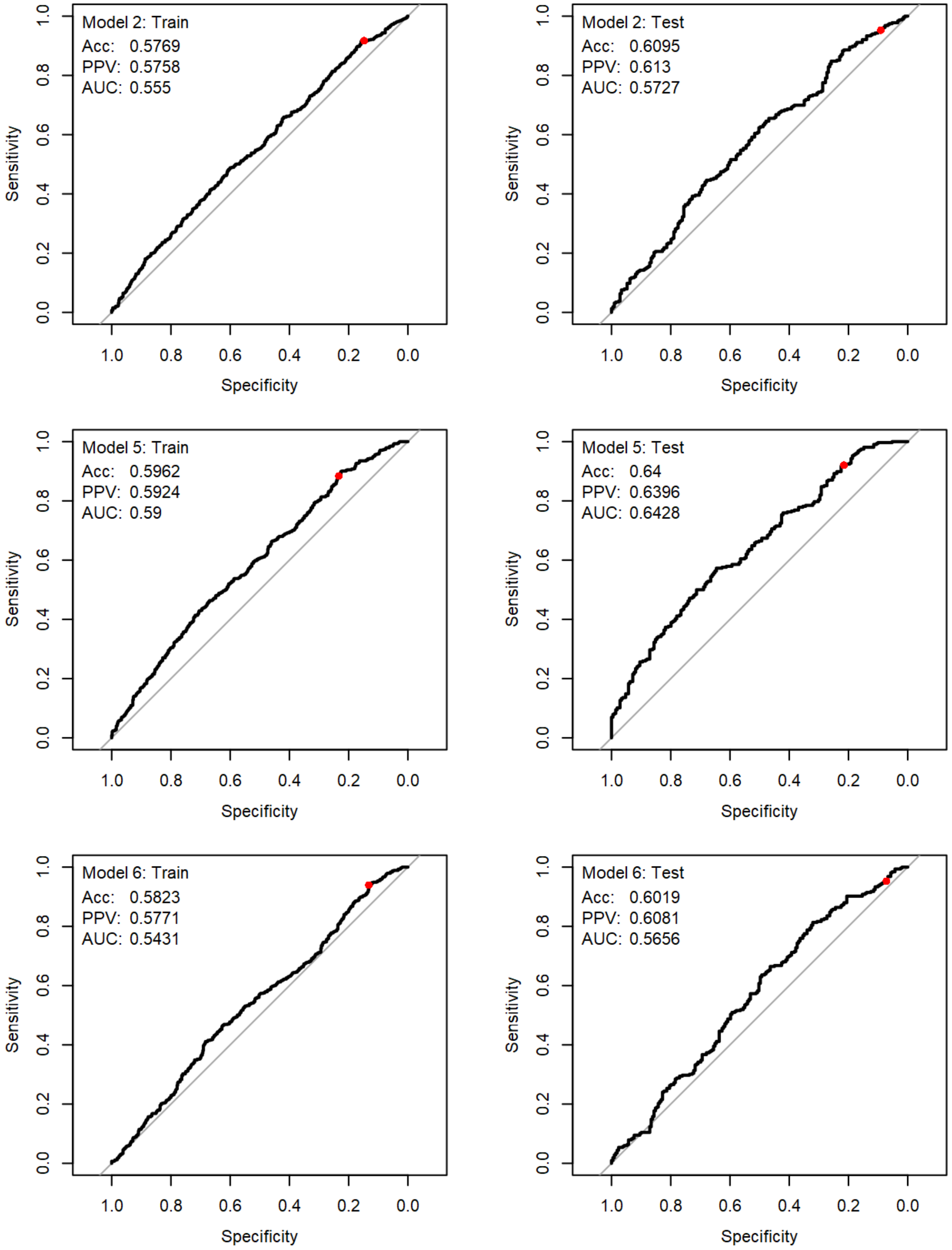


Figure 20: ROC curves for the logistic regression models selected by maximising PPV. Accuracy, positive predictive value (PPV), and area under the curve (AUC) are listed for each plot; the red dot represents model sensitivity and specificity.

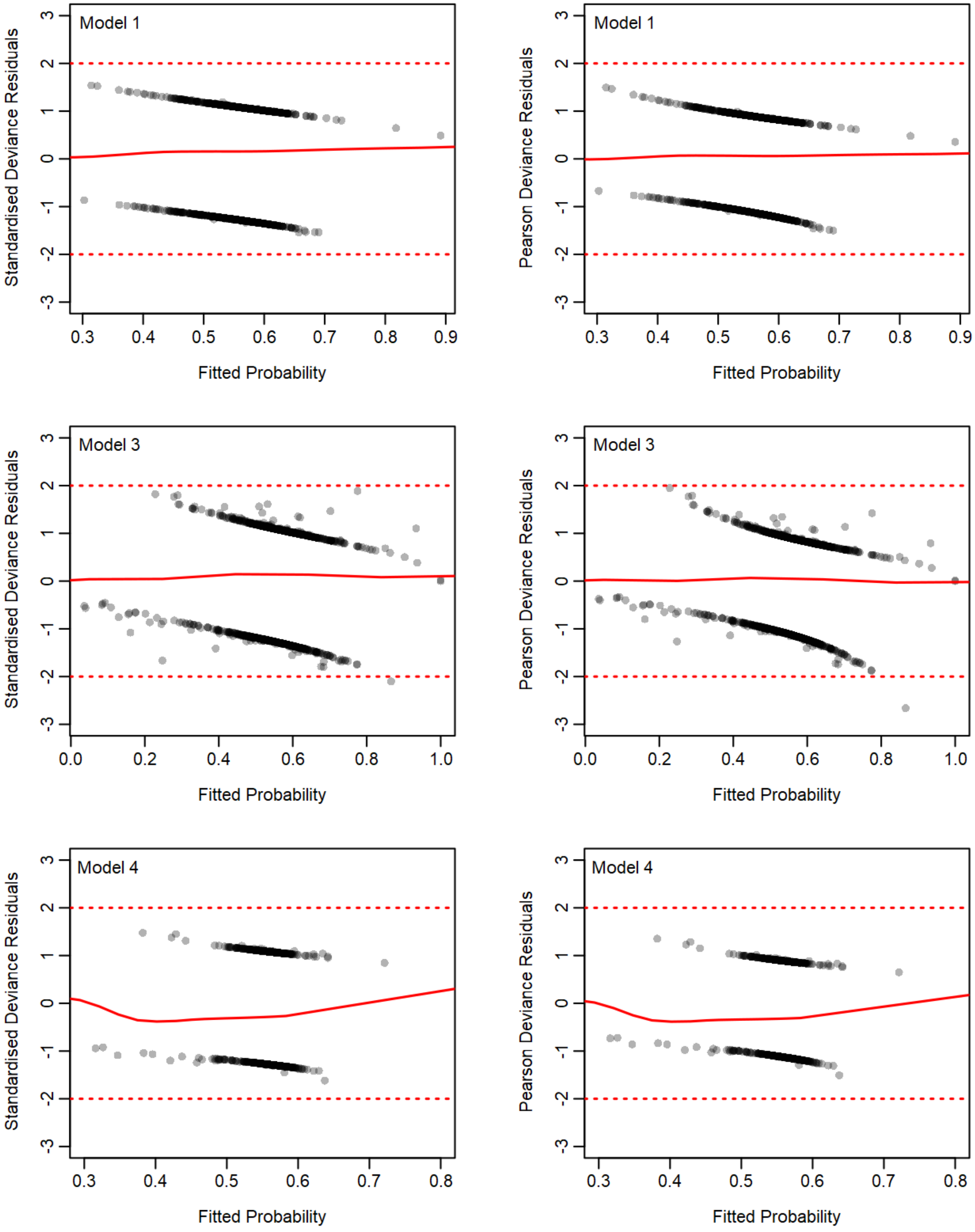


Figure 21: Standardised deviance residuals (left) and standardised Pearson residuals (right) for logistic regression models selected using accuracy. Solid red lines represent a loess fit to the residuals and should be approximately horizontal at zero; dotted red lines represent ± 2 standard deviations on the residuals.

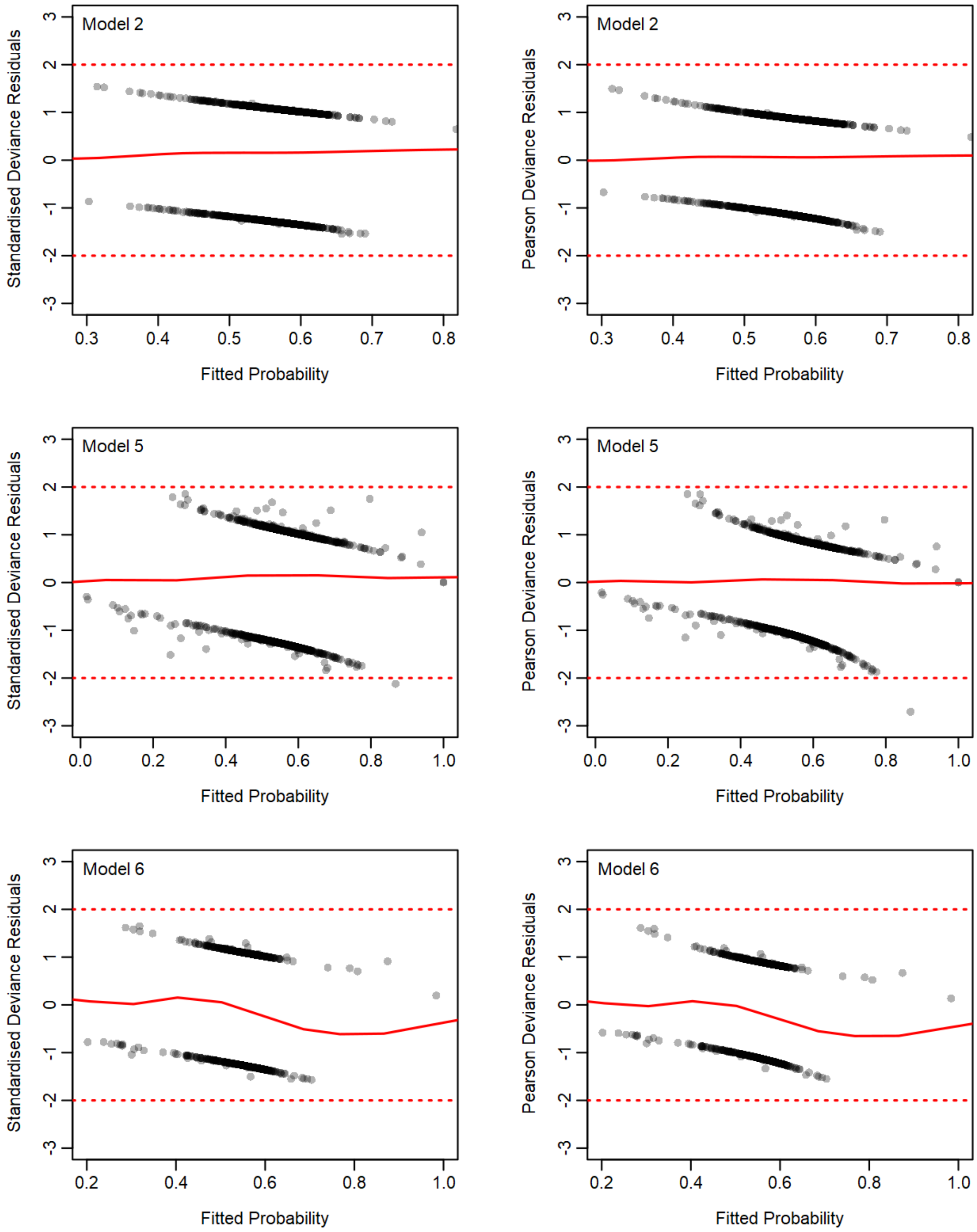


Figure 22: Standardised deviance residuals (left) and standardised Pearson residuals (right) for logistic regression models selected using PPV. Solid red lines represent a loess fit to the residuals and should be approximately horizontal at zero; dotted red lines represent ± 2 standard deviations on the residuals.

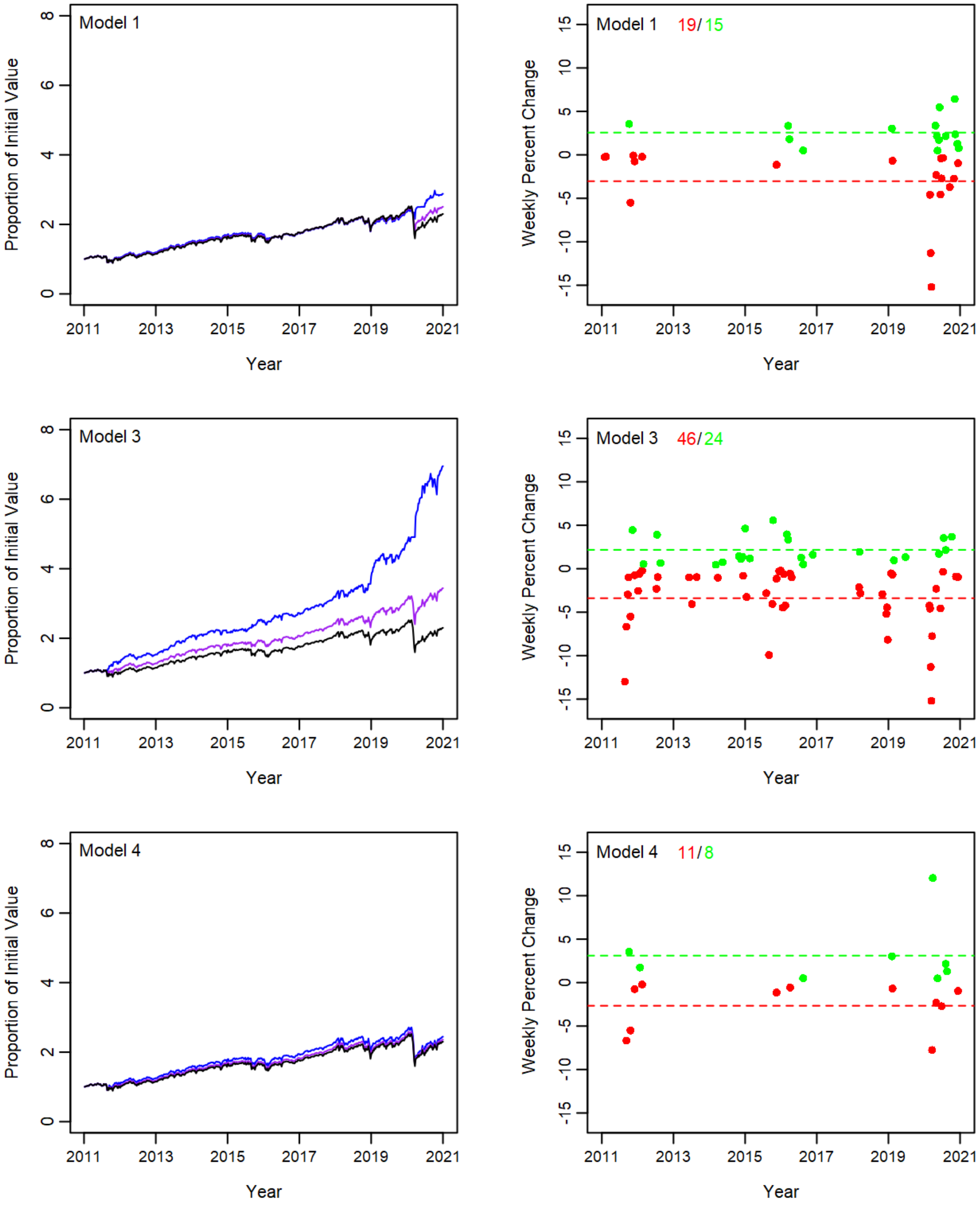


Figure 23: Relative value of the S&P 500 from 2011-2021 (left) and percent change for weeks in 2011-2021 for which an increase was not predicted (right) based on logistic regression models fit to the 1986-2010 data; the three models here were selected using overall accuracy. The black line represents a buy-and-hold strategy, while the blue line represents a strategy of selling all assets before a predicted decrease and then purchasing them back the next time there is a predicted increase. The purple line represents a less risky strategy of only selling 35% of assets before a predicted decrease and then purchasing them back the next time there is a predicted increase.

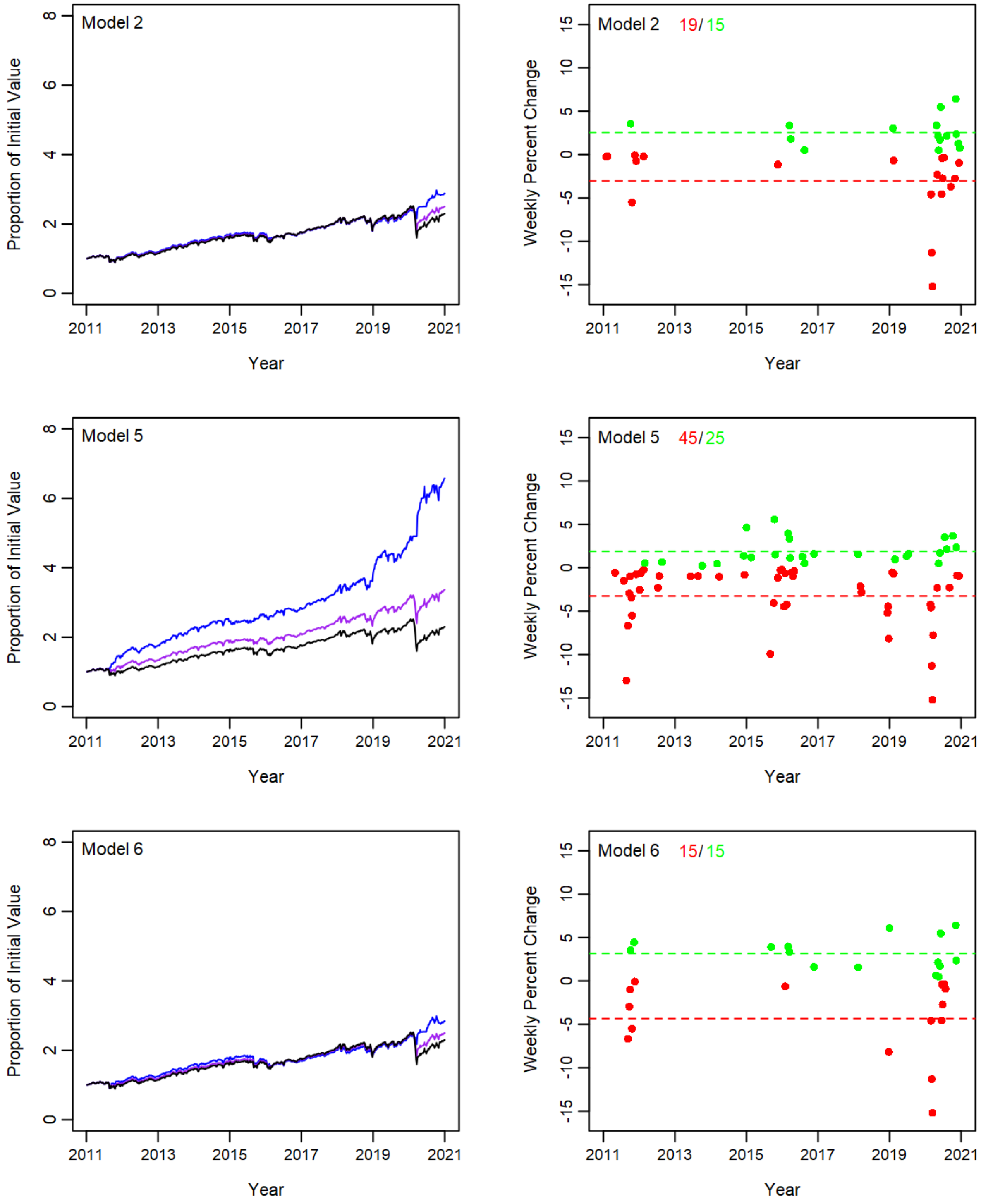


Figure 24: Relative value of the S&P 500 from 2011-2021 (left) and percent change for weeks in 2011-2021 for which an increase was not predicted (right) based on logistic regression models fit to the 1986-2010 data; the three models here were selected using PPV. The black line represents a buy-and-hold strategy, while the blue line represents a strategy of selling all assets before a predicted decrease and then purchasing them back the next time there is a predicted increase. The purple line represents a less risky strategy of only selling 35% of assets before a predicted decrease and then purchasing them back the next time there is a predicted increase.

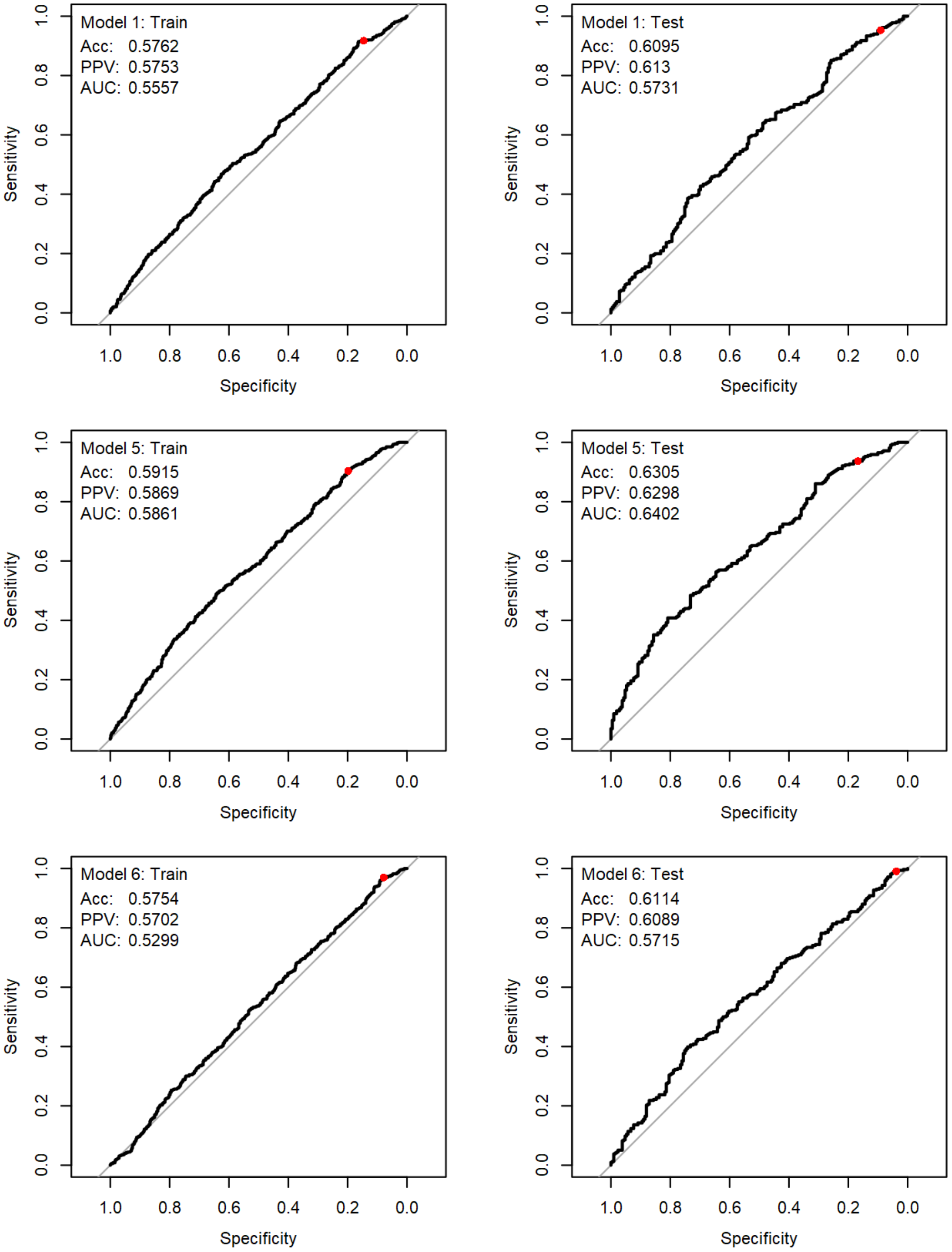


Figure 25: ROC curves for the LDA models selected by maximising accuracy. Accuracy, positive predictive value (PPV), and area under the curve (AUC) are listed for each plot; the red dot represents model sensitivity and specificity.

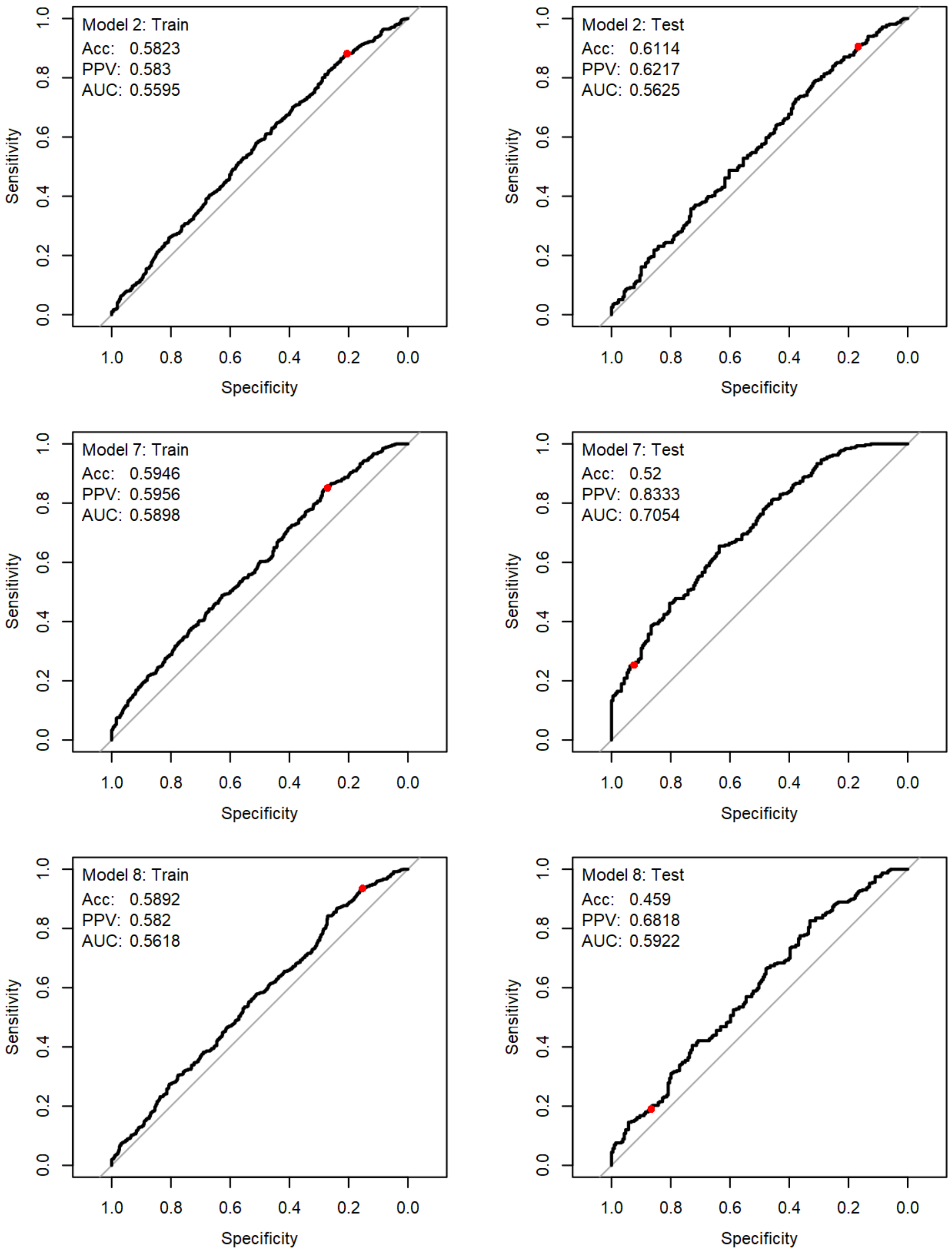


Figure 26: ROC curves for the QDA models selected by maximising accuracy. Accuracy, positive predictive value (PPV), and area under the curve (AUC) are listed for each plot; the red dot represents model sensitivity and specificity.

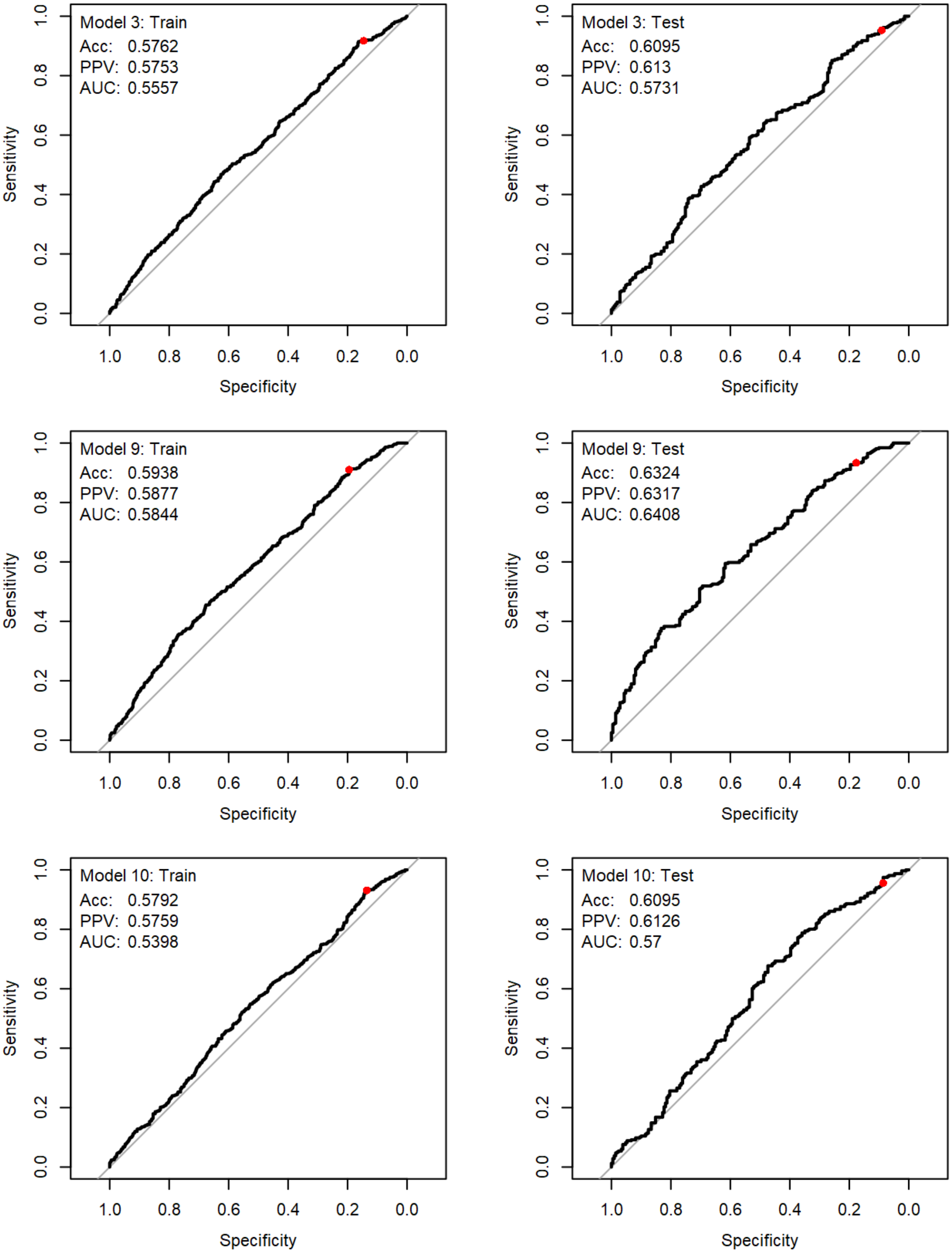


Figure 27: ROC curves for the LDA models selected by maximising PPV. Accuracy, positive predictive value (PPV), and area under the curve (AUC) are listed for each plot; the red dot represents model sensitivity and specificity.

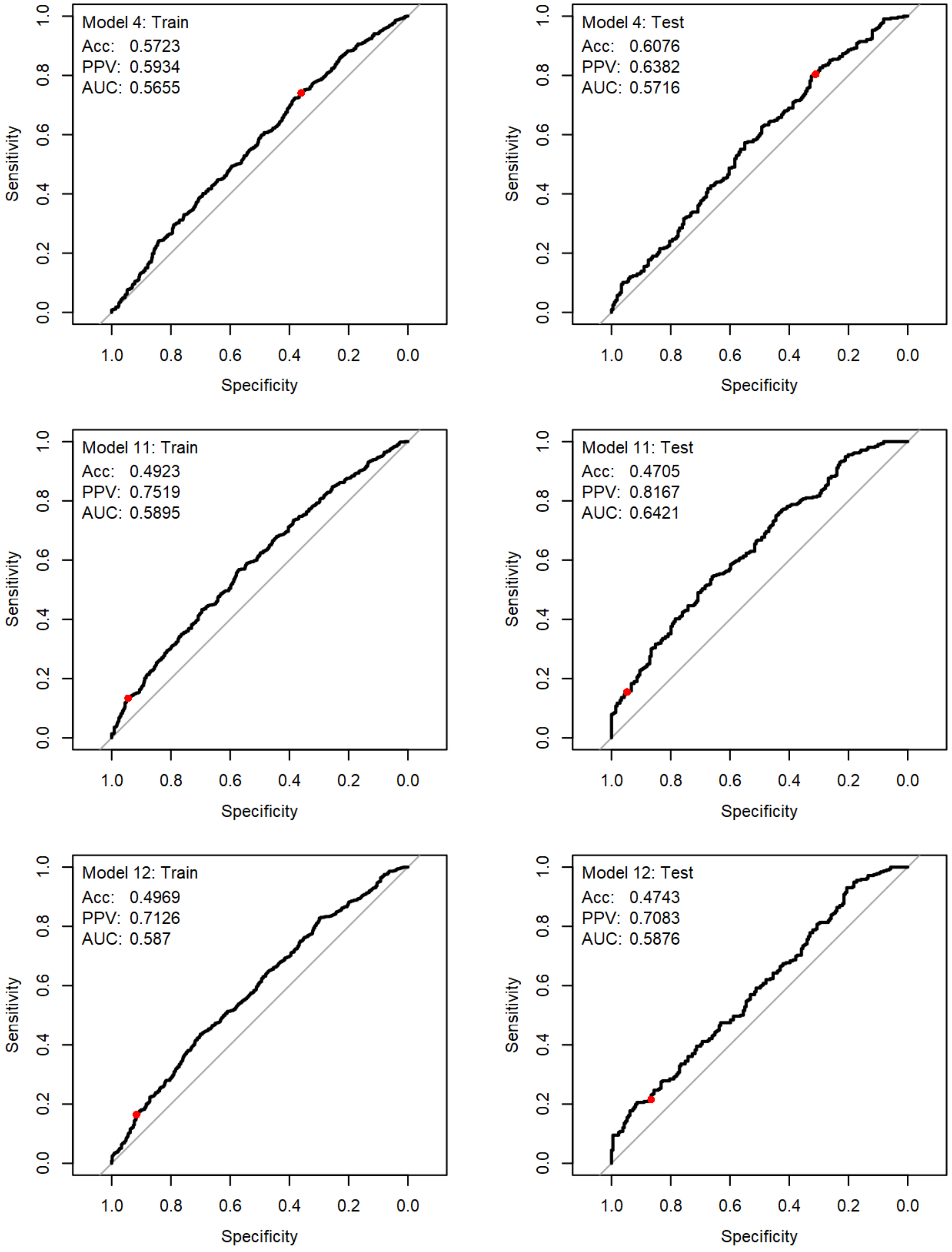


Figure 28: ROC curves for the QDA models selected by maximising PPV. Accuracy, positive predictive value (PPV), and area under the curve (AUC) are listed for each plot; the red dot represents model sensitivity and specificity.

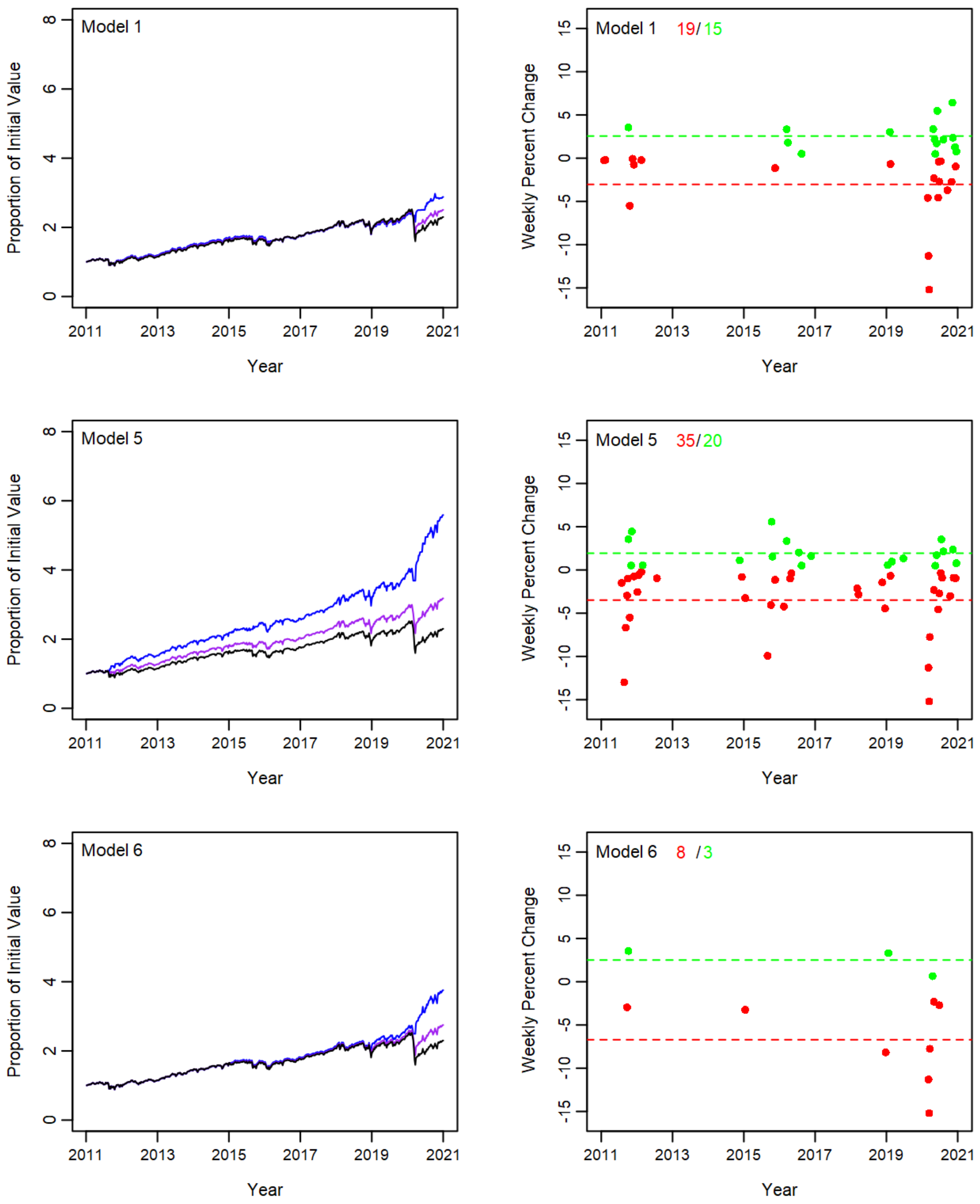


Figure 29: Relative value of the S&P 500 from 2011-2021 (left) and percent change for weeks in 2011-2021 for which an increase was not predicted (right) based on LDA models fit to the 1986-2010 data; the three models here were selected using accuracy. The black line represents a buy-and-hold strategy, while the blue line represents a strategy of selling all assets before a predicted decrease and then purchasing them back the next time there is a predicted increase. The purple line represents a less risky strategy of only selling 35% of assets before a predicted decrease and then purchasing them back the next time there is a predicted increase.

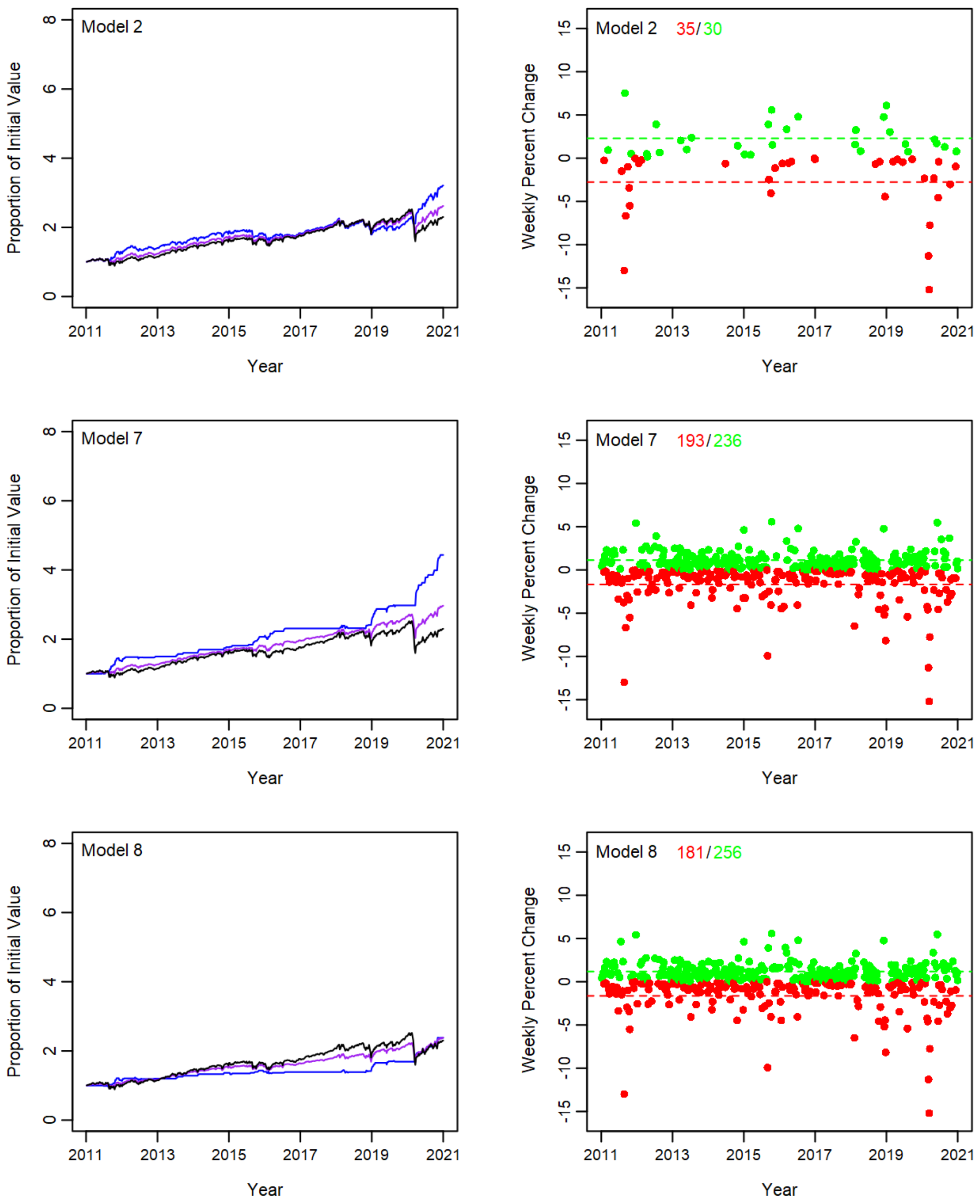


Figure 30: Relative value of the S&P 500 from 2011-2021 (left) and percent change for weeks in 2011-2021 for which an increase was not predicted (right) based on QDA models fit to the 1986-2010 data; the three models here were selected using accuracy. The black line represents a buy-and-hold strategy, while the blue line represents a strategy of selling all assets before a predicted decrease and then purchasing them back the next time there is a predicted increase. The purple line represents a less risky strategy of only selling 35% of assets before a predicted decrease and then purchasing them back the next time there is a predicted increase.

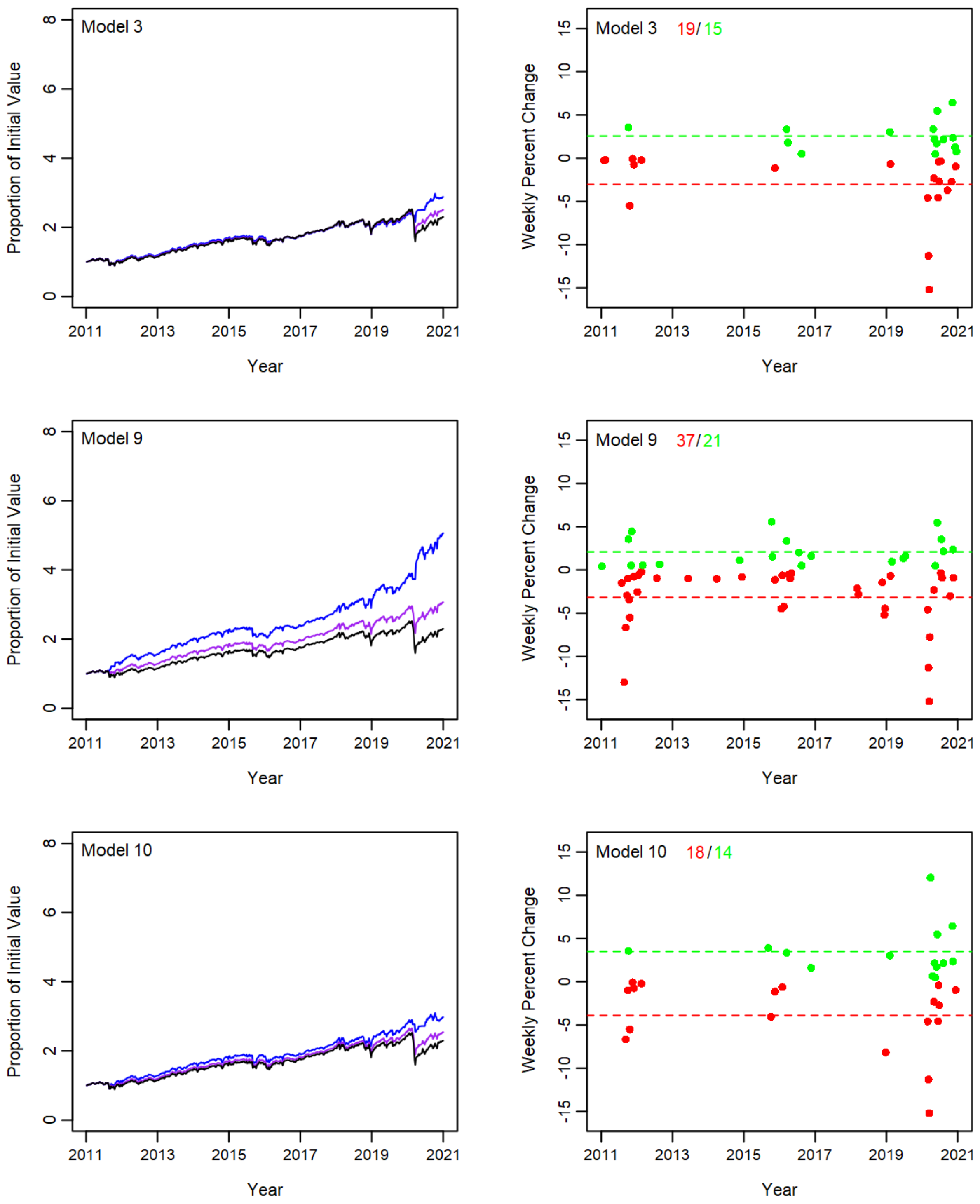


Figure 31: Relative value of the S&P 500 from 2011-2021 (left) and percent change for weeks in 2011-2021 for which an increase was not predicted (right) based on LDA models fit to the 1986-2010 data; the three models here were selected using PPV. The black line represents a buy-and-hold strategy, while the blue line represents a strategy of selling all assets before a predicted decrease and then purchasing them back the next time there is a predicted increase. The purple line represents a less risky strategy of only selling 35% of assets before a predicted decrease and then purchasing them back the next time there is a predicted increase.

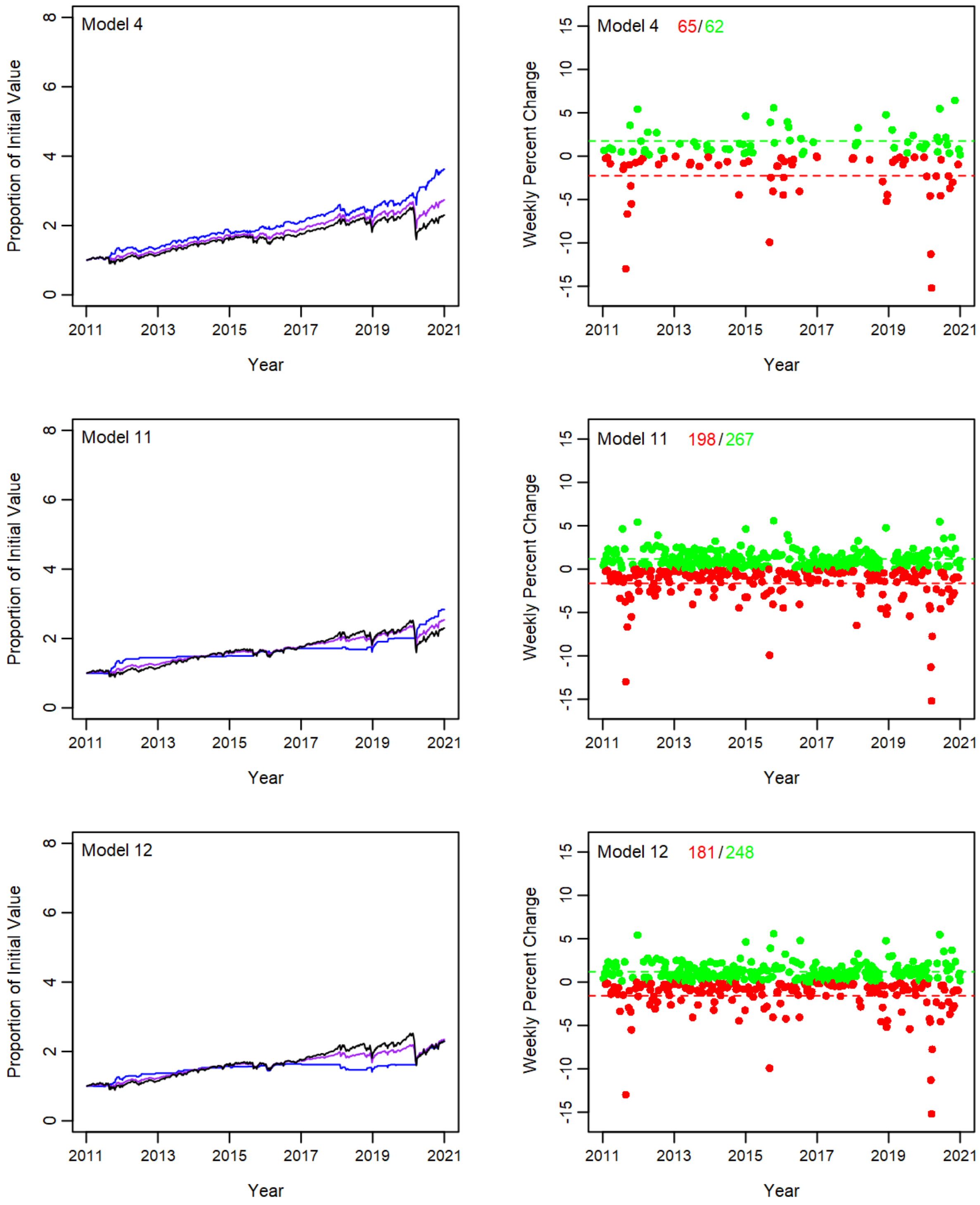


Figure 32: Relative value of the S&P 500 from 2011-2021 (left) and percent change for weeks in 2011-2021 for which an increase was not predicted (right) based on QDA models fit to the 1986-2010 data; the three models here were selected using PPV. The black line represents a buy-and-hold strategy, while the blue line represents a strategy of selling all assets before a predicted decrease and then purchasing them back the next time there is a predicted increase. The purple line represents a less risky strategy of only selling 35% of assets before a predicted decrease and then purchasing them back the next time there is a predicted increase.