Data Analytics and Machine Learning | Prof. Lochstoer

Problem Set 4

Using the StockRetAcct_sample.dta Stata dataset available at CCLE (Week 1), we will in this exercise code up a "machine learning" version of characteristics-based investment. That is, we will program an automated learning routine that finds the model specification -- which variables to include and the functional form these variables take -- and evaluate the out-of-sample performance of this model.

As described in the lecture notes for Topic 4, we will use the Elastic Net procedure, which includes Ridge Regressions and LASSO as special cases. The setup for this homework is given in slides 46 - 51 in Topic 4. The core idea is to develop a model where we estimate the portfolio weights of the MVE portfolio directly (as opposed to separately estimate the conditional mean returns and covariance matrix).

Question 1: Automatic Stock Picking Algorithm

a. Download the data. The firm-level characteristics you will use are lnIssue, lnProf, lnInv, and lnME. For each of these four characteristics, create new, additional characteristics as the squared value of the original characteristic. Name the new characteristics the same as the orignal, but with a "2" at the end. For instance, for lnProf, the squared value should be lnProf2. Further, create additional characteristics by multiplying each characteristic with lnME (except for lnME itself, which you already have squared). To name these, add _ME at the end. Thus, lnProf interacted with lnME is named lnProf_ME. You should have now gone from 4 to 11 characteristics.

(i) For each year in the sample, cross-sectionally demean each of the 11 characteristics. That is, for each characteristic and each year subtract the average value of that characteristic across stocks. Then add as final characteristic a column of 1's to the dataset. This effectively inserts an intercept in the relation between the MVE portfolio weight and the characteristics.

Next calculate the factor portfolio returns corresponding to each of these 12 characteristics, as explained on slide 48 in the Topic 4 note (F is implicitly defined in the equation there). Note that the factor corresponding to the constant is simply an equal-weighted portfolio of all stocks (the "market"). The overall idea is that with this approach you have a market factor and long-short characteristics factors.

Calculate and report the factor sample means and sample variance-covariance matrix for these 12 factors’ returns, as well as the factors' sample Sharpe ratios.

(ii) Next, you are to use the Elastic Net procedure (alpha = 0.5 in glmnet) to estimate the b coefficients. We will use the 25 year sample from 1980-2004 for the cross-validation exercise, and then we will use the 2005-2014 period for the out of sample testing. Here, we cannot use the pre-programmed cross-validation procedure in cv.glmnet. The reason is that the right- and left-hand side variables depend on the sample in a way not accounted for in the canned glmnet procedure.

You are to run a cross-sectional regression of average returns to the factors on the covariances of each factor with itself and the other factors (as in the bottom equation on slide 49 in Topic 4). A 5-fold cross-validation would tell you to first find the sample factor averages and sample factor covariance matrix in a 20-year subperiod, and then see how well the estimated *b* coefficients do in the 5-year out of sample period. In the out of sample period, the average returns are the 5-year average factor returns for this period and the covariances are the 5-year covariances in this period. Thus, due to the combination of time series info (average returns and sample covariance matrix) and the cross-sectional regression, our setting is a little more complicated than the standard cv.glmnet code.

So, to be clear: first, find sample average factor returns and covariance matrix from 1980-1999. Estimate the Elastic Net using the glmnet procedure (use family = 'Gaussian', alpha = 0.5). This gives you a matrix of b coefficients as a function of lambda. For each of these sets of b coefficients (each vector of b's corresponds to a particular lambda), calculate the mean squared error in the cross-validation out-of-sample period 2000-2004. Now you have MSE as a function of lambda for one 5-year fold. Then repeat using as in-sample data the 1980-1984 and 1990-2004 period. The out of sample data is then the 1985-1989 period. Get the MSEs as a function of lambda and save. Repeat until you have done all 5 folds. Then take the average MSE for each value of lambda. Pick the lambda that gives the smallest average MSE. Finally, estimate the elastic net on the full 1980-2004 sample period. Pick the *b*-coefficient that corresponds to the value of lambda you have chosen.

(iii) With the final *b*-vector in hand, calculate the out-of-sample average return, standard deviation, and Sharpe ratio for the corresponding estimated "ex ante" MVE portfolio with return b'F_t in the period 2005-2014.

(iv) Plot the cumulative return on this portfolio relative to that on the market (get market return using the value-weights in the sample, MEwt) over the 2005-2014 period, where you normalize the "MVE" portfolio's standard deviation to be the same as the market over this period. Compare. Note that one should really redo the estimation each year to get proper out of sample results that would mimic what you would do in the real world. Also, you could experiment in the in-sample cross-validation with different values for alpha to see what works best.