Master Thesis
LASSO-based predictive regression for stock returns

Hieu Hoang

December 10, 2019

Abstract

This is a place holder.

Table of Contents

Contents

1 Introduction

Prediction of stock returns has always been an important subject in finance, since an accurate prediction can help investors decide the portions of safe and risky assets in their portfolios, generating optimal wealth for their clients and themselves. As a result, a large volume of literature has been dedicated to developing asset pricing theory and predictive statistical models. One of the most popular and long-standing bases in the realm of asset pricing is the efficient market hypothesis (EMH), summarized and popularized in Fama and Malkiel (1970). According to EMH, equity premium is constant and reflects all available information. Hence, market anomalies and historical average of premia are analyzed to forecast asset returns.

On the other hand, financial econometricians focus on including relevant lagged financial and macroeconomic variables as predictors for equity premium as a way to exploit market inefficiency. Fama and French (1988), Schiller and Campbell (1998), among others explored the power of using valuation ratios, e.g. dividend-price ratio, dividend yield, earnings-price, book-to-market ratio to forecast long-term returns of stock, while Fama (1990), Schwert (1990), and related papers showed correlation between bonds (treasury and corporate) and stock returns. As a result, a healthy number of literature demonstrated and cemented this idea, see, among Hodrick (1992), Kothari and Shanken (1997), Lamont (1998), Pontiff and Schall (1998).

As of the late 1990s and early 2000s, the common consensus within the field is that excess stock returns can be predicted (Welch and Goyal, 2008). However, along with new tools come new challenges. One of which is that the results of aforementioned findings may very well be spurious. (Stambaugh, 1999) showed that when the innovation of a predictor is correlated with excess return, which is the case for many valuation ratios, the resulting estimated coefficient is biased and exhibits sharply different finite-sample properties from the standard case. One apparent example is the dividend yield which has the same return component with equity premium. A second source for spurious result stems from the high persistence of predictors. It has been well-known that a regression model with integrated or near-integrated predictors may sometimes produce 'non-sense correlation' where highly significant betas and high R-squared values are obtained while no "real" and meaningful correlation exists, except for the case of co-integrated series; see among Yule (1926), Granger and Newbold (1974), Phillips (1986), Granger and Newbold (2001), and Engle and Granger (1987). Last but not least, the problem of near-collinear predictors is also apparent for financial econometricians. As stated above, many highly persistent time series variables exhibit meaningless high cross-correlation scenario, but high correlation nonetheless. Furthermore, some ratios are by construction derived from other ratios or macroeconomic variables, hence the possible long-term co-movement between said ratios. This in turn causes the design matrix to become nearly singular (asymptotically singular), which produce estimation inconsistencies and failures in central limit theory in least square regression even in the case of stationary predictors and strong regression signal (Phillips, 2016), much less for our case of high-persistence.

For the reasons stated above, OLS may not be the best way to go. The late 1990s and early 2000s also witnessed a relatively new method for estimating linear models: the lasso (Tibshirani, 1996). Instead of just minimizing the residual sum of squares, the lasso further applies a constraint to the sum of absolute values of estimated coefficients. This penalty term helps shrink coefficient estimates and at the same time encourages some variables to take on zero as coefficient, effectively eliminate them from the model. Thus, the lasso has the advantages of both (continuous) model selection and variance reduction by trading off some amount of bias, which may increase predictive performance. Belonging to the penalized least squared family of regression method, lasso also benefits from stability when there exist high collinearity between predictors by preventing coefficient inflation as in the case of its brother shrinkage variation, the ridge regression. Still, lasso suffers from a number of problems, mainly related to inference. Additionally, we still have the problem of near-singular design matrix and near-integrated and integrated series, potentially cointegrated in our hand.

This paper will be organized as followed. This section gave an introduction to the paper. Next, I will review some literature regarding the problem of asymptotically degenerate design matrix, and how it makes OLS estimations invalid. An overview of lasso (and its variants) and how lasso can be used to combat our prevailing problems will also be discussed in this section. The third section will be about the technical details of Adaptive Lasso (henceforth alasso), with focus on the inference of coefficient estimates. The fourth section compares predictive performance of alasso with regular lasso, autoregressive of order 1, and OLS in simulation settings. Lastly, I will apply alasso to Goyal's data set used in Welch and Goyal (2008), with updates until 2018, to assess its real-world predictive performance. Conclusion and extension will be given as closing thoughts.

2 Literature Review

In this section, I will first review a number of the literature that discusses our three big problems in the context of OLS regression: correlation between innovation of lagged predictor and regression disturbance; mixed roots, high persistence predictors; and near-singular design matrix. Next, we will take a look at how lasso-type regression, specifically alasso, can help alleviate parts of our problems.

OLS, or ordinary least squares, is a long-standing powerhouse in the scene of linear regression. The objective of OLS is to find a set of coefficients that minimizes the squared differences between observed dependent variables and its predicted value. Thank to its readily available analytical solution, fast computation, and well-studied inference, it is widely used in both cross-sectional and time series data alike. However, there are a set of assumptions required to make the OLS estimates valid. — Specifically, the stochastic processes involved must be stationary and ergodic. — more precise here. reformulate, validity is not necessarily restricted to stationarity and ergodicity, more important is that the disturbances are uncorrelated to the regressors. — However, evidences for stationarity of valuation ratios are mixed and shaky. Roll (2002) argues that under rational expectation, asset price is non-stationary due to its dependence on expectation of future quantities. Yet, metrics that are constructed as functions of price, e.g market-to-book, earning-price ratios, dividend yields may exhibit different root characteristics (Phillips, 2015). At the same time, most remaining series show high yet imprecisely determined degree of persistence, leading to the problem of mixed roots, possibly cointegrated, regression. Phillips (2015) also discussed 'misbalancing' issue, where predicted variable and predictors have different memory type. The solution out is not straight-forward. Elliott and Stock (1994) discussed two common simple solutions: ignore the problem altogether, or determine the post-regression inference by pretesting predictors for unit roots. Both lead to the substantial over-rejection of the null of no significance. The solution proposed in the same paper involves Bayesian statistic, thus may not be appealing to some. In another approach, a local-to-unity autoregressive specification in the form of $\rho = 1 + \frac{c}{n}$ is used to conduct asymptotic theory. However, the introduction of the unknown parameter brings more issues. Since localizing coefficient c is not consistently estimable, asymptotic bias cannot be corrected, leading to nonstandard limit theory. Phillips and Lee (2013) discussed c and suggested possible solutions.

Another violation of OLS assumptions comes in the form of high to perfect correlation between predictors. In the case of perfect correlation, removing appropriate regressor is a common remedy. When the correlation is not perfect, determinant(s) of design matrix gets into the vicinity of zero, causing computational difficulty in matrix inversion and inflated coefficient estimates. However, removal of regressors is not always an preferable option since each regressor may contain some additional information that can improve the model fit or prediction. In the case of predicting excess returns, some predictors may contain information about market inefficiency despite high correlation due to a common variable in their construction. This kind of construction also leads to possible co-movement of predictors, causing singularity in the limit (near-singular design matrix). As variable frequency increases such as in the case of financial data, singularity can come very quickly.

In this paper, I would like to introduce a relatively new method of estimation for linear model that has the ability to hopefully overcome some of the aforementioned issues. Proposed and discussed by Tibshirani (1996), the lasso (least absolute shrinkage and selection operator) exhibits some more preferable properties than the well-known OLS.

First, while shrinking coefficients introduces some amount of bias, it helps to prevent estimate inflation in presence of high-collinearity. In fact, for the case of near-singular design, the lasso estimates are consistent, and with an appropriate choice for shrinkage parameter λ, limiting distribution is normal (Knight et al., 2000; Knight, 2008). This result is especially handy for the case in this paper.

Second, lasso can set some coefficients to zero, effectively performs continuous model selection. Via this mechanism, variance is reduced and hence accuracy may increase in the case of predictive regression. Continuous model selection has some advantages over discrete model selection, for example subset selection, where small change in data can lead to substantially different selection outcome, or be trapped in a local optimum (Breiman, 1995). Furthermore, as the number of regressors increases, discrete selection is computationally hard. On the other hand, continuous selection process is more stable, intepretable, and can scale easily with a large number of variables (Tibshirani, 1996). However, plain-vanilla lasso is not always consistent in identifying the right subset of variables, and does not always exhibit 'oracle properties' (consistency in variable selection and asymptotic normality) (Meinshausen and Buhlmann, 2004; Zou, 2006). Hence, the adaptive lasso is proposed by Zou (2006). Alasso assigns weights to each coefficients, and if the weights are cleverly chosen and data-driven, alasso can enjoys oracle properties.

Last but not least, the alasso works well in the case of mixed degree of persistence in predictive regression mentioned above. It can even adapt to system of predictors that exhibits cointegration by assigning appropriate penalty level inside the system without knowing the identity of these predictors. Lee et al. (2018) establishes and demonstrates a simple condition on λ that leads alasso to oracle properties without knowledge of persistence level in advance.

With all the favorable theory at hand, I will embark on testing the performance of alasso in both simulation and real data settings. But first, I will re-establish important results mentioned above in a more concrete manner.

3 The Adaptive Lasso and Its Advantages

In this section, we will discuss the advantage of alasso. The theoretical framework will closely follow Tibshirani (1996); Knight (2008); Zou (2006); Lee et al. (2018).

THE MODEL The linear model is assumed to be as following, adapted from Lee et al. (2018):

$$\begin{aligned} y_i &= \sum_{l=1}^{p_z} z_{il}\alpha_l^* + \sum_{l=1}^{p_x} x_{il}\beta_l^* + \sum_{l=1}^{p_c} x_{il}^c\phi_l^* + \varepsilon_i \\ &= \boldsymbol{z}_i'\alpha^* + \boldsymbol{x}_i'\beta^* + \boldsymbol{x}_i'^c\phi^* + \varepsilon_i \\ &= \boldsymbol{w_i'\theta} + \varepsilon_i, \\ \boldsymbol{y} &= \boldsymbol{w'\theta} + \boldsymbol{\varepsilon}. \end{aligned} \tag{1}$$

for $i = 1, \ldots, n$, where $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{ip_z})'$, $\boldsymbol{x}_i^c = (x_{i1}^c, \ldots, x_{ip_z}^c)'$, and $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip_x})'$ represent the stationary, cointegrated, and unit root regressors, respectively, and $p = p_z + p_c + p_x$ is the total number of regressors in the model. $\boldsymbol{w_i}$ is $p \times 1$ vector of all regressors; $\boldsymbol{\theta}$ is $p \times 1$ vector of all coefficients. The presence of heterogeneous degrees of persistence and cointegration in our regressors nicely follows practical situation in predicting excess return with multiple valuation ratios and macroeconomic variables. — add definition for stacked matrix as well —

One deviation from the model laid out in Lee et al. (2018) is that regressors are allowed to have increasingly strong correlation between one another. Specifically, as sample size n increases, the degree of correlation between said regressors also rises. This setting is included to emulate near-singular design matrix phenomenon frequently encountered in stock return predictive regression. Later, we will see that the sequence of tuning parameter λ proposed by Lee et al. (2018) intended for mixed root in regressors is also helpful in combating near-singular design. Formally, define matrix C_n as in Knight (2008):

$$C_n = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{w}_i \boldsymbol{w}_i' \tag{2}$$

that is nonsingular for each n except for when $n \to \infty$, then

$$C_n \to C, \tag{3}$$

where C is singular. In practice, near-singular design can be detected by checking whether the smallest eigenvalue of C_n is small in comparison to its trace. For some sequence $\{a_n\}$ tending to infinity, we assume that

$$a_n(C_n - C) \to D_0 \tag{4}$$

where D_0 is positive definite on the null space of C, that is $\boldsymbol{v}' D_0 \boldsymbol{v} > 0$ for non-zero $\boldsymbol{v}$ with $C\boldsymbol{v} = 0$. Another assumption is that near-singularity affect all regressors in our model.

To ensure stationarity of y_i, the effect of non-stationary regressors (including unit root components in cointegrated ones) are kept small using local-to-zero coefficients. This type of coefficient design is also useful to model the weak signal-to-noise ratio in predictive regression (Phillips and Lee, 2013; Lee et al., 2018). The $p \times 1$ true coefficient $\theta_n^* = (\theta_{jn}^* = \theta_j^{0*}/n^{\delta_j})_{j=1}^p$ where $\theta_j^{0*} \in \mathbb{R}$ is a fixed constant independent of sample size, and $\delta_j \in [0, 1)$. In cases where $\theta_j^{0*} = 0$, δ_j is also set to zero. θ_n^* thus gets smaller with increasing sample size for $\theta_j^{0*} \neq 0$ and $\delta_j \in (0, 1)$, and approach zero as n tends to infinity.

— More definition about $\bar{\delta}$, set of true parameters M^*, set of screened parameters $\hat{M}_n$ —

On the front of identification, assumptions about error terms must also be made. While correlation between regression error and the innovation of non-stationary regressors $\boldsymbol{x}_i$ is allowed, correlation between regression errors and innovation of stationary and cointegrated regressors is excluded (see **Assumption 3.1** and **Remark 3.1**, Lee et al. (2018)).

THE LASSO Lasso is a technique for estimation of linear models that utilize regularization in order to shrink coefficients and perform variable selection at the same time (Tibshirani, 1996). Lasso objective function for model (1) are defined to be

$$\sum_{i=1}^{n}(y_i - \boldsymbol{w}_i'\boldsymbol{\theta})^2 + \lambda_n \sum_{j=1}^{p} |\theta_j|, \tag{5}$$

which is essentially least square with an additional $\ell 1$ penalty that helps force estimates of "small" parameter towards zero. Despite of course introducing some bias into the estimates, lasso may reduce estimation variance. In the limit, truly "small" parameters are zero with probability tends to 1 while all others are discernibly not zero. Therefore, in cases where true parameters are zero, no biases are produced and variance is reduced, a win-win situation (superefficiency, as termed in Knight (2008)). On the other hand, such regularization causes bias in estimates for non-zero true coefficients while typically does not improve estimation variance considerably. Naturally, we want estimators that achieve superefficiency when $\theta_j^* = 0$ and produce no asymptotic bias otherwise. Studying how estimators behave in the limit regarding the choice of tuning parameter λ is hence a great way to come up with a desirable one. Knight et al. (2000) find that if λ is treated as a sequence dependent on sample size, and design matrix is non-singular, $O(n)$ growth rate of $\{\lambda_n\}$ is sufficient to obtain $\sqrt{n}$-consistency and non-degenerate limiting distribution. However, in this paper we are more interested in cases where the design matrix is near-singular.

NEAR-SINGULAR DESIGN MATRIX Assume model (1) with C_n satisfies (2), (3), and (4), that C is singular and D_0 is positive definite in the null space of C. Define $b_n = (n/a_n)^{1/2}$ for $\{a_n\}$ satisfies (4), and define Z_n to be

$$Z_n(\boldsymbol{u}) = \sum_{i=1}^{n}[(\varepsilon_i - \boldsymbol{u}'\boldsymbol{w}_i/b_n)^2 - \varepsilon_i^2] + \lambda_n \sum_{j=1}^{p}(|\theta_j + u_j/b_n| - |\theta_j|). \tag{6}$$

This equation is a rescaled version of the objective function in (5) with constants subtracted so that convergence is ensured. If $\hat{\boldsymbol{\theta}}$ minimizes (5) then $b_n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ minimizes (6). The following theorem is adapted from Knight et al. (2000):

Theorem 1 (adapted from Theorem 1, Knight (2008)) *Define* $\boldsymbol{\Omega}$ *to be a zero mean multivariate normal random vector such that* $\mathrm{Var}(\boldsymbol{u}'\boldsymbol{\Omega}) = \sigma^2\boldsymbol{u}'D_0\boldsymbol{u}$ *positive for each nonzero* $\boldsymbol{u}$ *that satisfies* $C\boldsymbol{u} = \mathbf{0}$*. Let* $\hat{\boldsymbol{\theta}}_n$ *minimizes* (5) *for* $\lambda_n \geq 0$*. If* $\lambda_n/b_n \to \lambda_0 \geq 0$ *then*

$$b_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{d} \arg\min\{Z(\boldsymbol{u}) : C\boldsymbol{u} = \mathbf{0}\},$$

where

$$Z(\boldsymbol{u}) = -2\boldsymbol{u}'\boldsymbol{W} + \boldsymbol{u}'D_0\boldsymbol{u} + \lambda_0\sum_{j=1}^{p}\{u_j\,\mathrm{sgn}(\theta_j) + |u_j|I(\theta_j = 0)\}.$$

This theorem reveals many interesting insights. With a proper choice of $\{\lambda_n\}$, that is the sequence converges to a finite non-negative value λ_0, lasso estimate has normal limiting distribution. In the case of $\lambda_0 = 0$, lasso, somewhat trivially, is consistent and has the same limiting distribution as the OLS does since the penalty term vanishes. Less trivial is the case where λ_0 takes on positive value. Suppose all true parameters $\theta_1, \ldots, \theta_p$ are non-zero, lasso estimate is unfortunately biased, and the amount of bias depends on the vector $\boldsymbol{u}$ and the signs of each element in the vector of true parameter $\boldsymbol{\theta}$. By assumption in Theorem 1, the null space of C is the space of vectors $\boldsymbol{u}$ with $u_1 + \ldots + u_p = 0$, therefore $Bias(\hat{\boldsymbol{\theta}}_n) = 0$ if true parameters all have the same sign. In another example case where $\theta_1 \neq 0$ and $\theta_2 = \ldots = \theta_p = 0$, the joint limiting distribution of $b_n(\hat{\boldsymbol{\theta}}_{nj} - \theta_j)$ for $p = 2, \ldots, p$ will have positive probability mass at $\mathbf{0}$. Since the limiting distribution lies in the null space of C as in (4), $b_n(\hat{\theta}_{n1} - \theta_1)$ is implied to have positive probability mass at 0 (Knight, 2008). This indicates possible asymptotic bias in the estimates of non-zero parameters. Another downside of singularity in the limit is the slower convergence rate of estimates to their limiting distributions due to $a_n \to \infty$ and $b_n = (n/a_n)^{1/2}$ hence b_n is only of $o(\sqrt{n})$. The exact margin however, still depends strongly on the growth rate of n_n.

MIXED ROOT Aside from near-singular design matrix, unknown degree of persistence also poses problems with not only OLS, but also with lasso. In the presence of possible unit roots (exact roots are unknown) in regressors, OLS estimates is bias in the limit due to serial dependence in the innovations. With lasso, the variable screening effect is very sensitive to the choice of tuning parameter, and each set of regressors is affected differently (**Corollary 3.7** and **Remark 3.8**, Lee et al. (2018)). As expected, when $\lambda_0 = 0$, lasso's limiting distribution collides with that of OLS since no selection takes place anymore. More interesting are cases where $\lambda_0 \neq 0$. For

$\lambda_0 \in (0, \infty)$, only the set of stationary receives selection. In this setting, λ_0 is still too small to have an effect on non-stationary part. Choosing the sequence to growth even faster, where $\lambda_n/\sqrt{n} \to \infty$ and $\lambda_n/n \to 0$, drags down convergence rate of estimates for the stationary set while still does not hit non-stationary set. Raising growth rate further however starts to introduce inconsistency (**Lemma 3**, Zou (2006)).

ADAPTIVE LASSO Introduced by Zou (2006), alasso is designed to overcome possible inconsistencies in variable selection of lasso. Alasso involves applying individual weight to the tuning parameter of each regressor. The objective function of alasso is defined as:

$$\sum_{i=1}^{n}(y_i - \boldsymbol{w}_i'\boldsymbol{\theta})^2 + \lambda_n \sum_{j=1}^{p} \hat{\tau}_j|\theta_j|, \tag{7}$$

where $\hat{\tau}_j$ is the individual weight for the corresponding regressor j. — what is the weight suggested? Why is it OLS? (lee) Present Oracle properties and condition for it Oracle properties in the case of mixed root Is the rate for oracle properties appropriate for near-singular design? —

4 Simulation study

5 Application: Goyal's data set

References

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics 37*(4), 373–384.

Elliott, G. and J. H. Stock (1994). Inference in time series regression when the order of integration of a regressor is unknown. *Econometric theory 10*(3-4), 672–700.

Engle, R. F. and C. W. Granger (1987). Co-integration and error correction: representation, estimation, and testing. *Econometrica: journal of the Econometric Society*, 251–276.

Fama, E. F. (1990). Stock returns, expected returns, and real activity. *The journal of finance 45*(4), 1089–1108.

Fama, E. F. and K. R. French (1988). Dividend yields and expected stock returns. *Journal of financial economics 22*(1), 3–25.

Fama, E. F. and B. G. Malkiel (1970). Efficient capital markets: A review of theory and empirical work. *The journal of Finance 25*(2), 383–417.

Granger, C. and P. Newbold (1974). Spurious regressions in econometrics. *Journal of Econometrics 2*(2), 111–120.

Granger, C. and P. Newbold (2001). Spurious regression in econometrics. *ECONOMETRIC SOCIETY MONOGRAPHS 33*, 109–118.

Hodrick, R. J. (1992). Dividend yields and expected stock returns: Alternative procedures for inference and measurement. *The Review of Financial Studies 5*(3), 357–386.

Knight, K. (2008). Shrinkage estimation for nearly singular designs. *Econometric Theory 24*(2), 323–337.

Knight, K., W. Fu, et al. (2000). Asymptotics for lasso-type estimators. *The Annals of statistics 28*(5), 1356–1378.

Kothari, S. P. and J. Shanken (1997). Book-to-market, dividend yield, and expected market returns: A time-series analysis. *Journal of Financial Economics 44*(2), 169–203.

Lamont, O. (1998). Earnings and expected returns. *The journal of Finance 53*(5), 1563–1587.

Lee, J. H., Z. Shi, and Z. Gao (2018). On lasso for predictive regression.

Meinshausen, N. and P. Buhlmann (2004). Consistent neighbourhood selection for sparse high-dimensional graphs with the lasso. Seminar für Statistik, Eidgenössische Technische Hochschule (ETH), Zürich.

Phillips, P. C. (1986). Understanding spurious regressions in econometrics. *Journal of econometrics 33*(3), 311–340.

Phillips, P. C. (2015). Halbert white jr. memorial jfec lecture: Pitfalls and possibilities in predictive regression. *Journal of Financial Econometrics 13*(3), 521–555.

Phillips, P. C. (2016). Inference in near-singular regression. In *Essays in Honor of Aman Ullah*, pp. 461–486. Emerald Group Publishing Limited.

Phillips, P. C. and J. H. Lee (2013). Predictive regression under various degrees of persistence and robust long-horizon regression. *Journal of Econometrics 177*(2), 250–264.

Pontiff, J. and L. D. Schall (1998). Book-to-market ratios as predictors of market returns. *Journal of Financial Economics 49*(2), 141–160.

Roll, R. (2002). Rational infinitely lived asset prices must be non-stationary. *Journal of banking & finance 26*(6), 1093–1097.

Schiller, J. and J. Campbell (1998). Valuation ratios and the long-run stock market outlook. *Journal of*.

Schwert, G. W. (1990). Stock returns and real activity: A century of evidence. *The Journal of Finance 45*(4), 1237–1257.

Stambaugh, R. F. (1999). Predictive regressions. *Journal of Financial Economics 54*(3), 375–421.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological) 58*(1), 267–288.

Welch, I. and A. Goyal (2008). A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies 21*(4), 1455–1508.

Yule, G. U. (1926). Why do we sometimes get nonsense-correlations between time-series?–a study in sampling and the nature of time-series. *Journal of the royal statistical society 89*(1), 1–63.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association 101*(476), 1418–1429.