Modeling the S&P 500 Index using the Kalman Filter and the LagLasso

Nicolas Mahler
ENS Cachan & UniverSud
CMLA UMR CNRS 8536
and
Telecom Paristech (TSI)
LTCI UMR Institut Telecom/CNRS 5141
nico.mahler@gmail.com

Abstract

This article introduces a method to predict upward and downward monthly variations of the S&P 500 index by using a pool of macro-economic and financial explicative variables. The method is based on the combination of a denoising step, performed by Kalman filtering, with a variable selection step, performed by a Lasso-type procedure. In particular, we propose an implementation of the Lasso method called LagLasso which includes selection of lags for individual factors. We provide promising backtesting results of the prediction model based on a naive trading rule.

1 Introduction

We consider the problem of predicting monthly movements of the S&P 500 index, and assume that a small subset of macro-economic and financial predictors can efficiently represent the exogenous influence on S&P 500. The influence of each of these predictors can change over time and it can be lagged. Additionnally, according to economists (cf.[8]), S&P 500 is sensitive to the variations of those predictors around their own trend rather than to the variations themselves. Therefore, we need to filtrate the predictors : a linear state-space model is first proposed for each of them and their innovation residuals are computed with the Kalman algorithm (cf.[4],[5],[6],[7]). These residuals are then used to predict S&P 500 variations : the most informative residuals are identified thanks to the Lasso method, a procedure which aims at minimizing a L2 regression fit under L1 penalty. This constraint allows for a sparse selection which is not only a gain in terms of interpretability, but which allows for variance reduction leading to more accurate predictions. The issue of lagged influence between variables is adressed by slightly modifying the Lasso. Indeed, as shown in [1] and [2], the Lasso is intimately connected to the LARS algorithm, which is an iterative procedure of variable selection generalizing the concept of bissector in a multidimensional framework. The basic idea consists in writing a variant of the Lasso, where both a variable and a lag are selected at each step, all the other lags of the variables being then eliminated from the possible further selections : we call it the LagLasso procedure. This approach is finally prospective, in the sense that we would like not only to build a competitive prediction method for S&P 500 but also to clearly state the problem of lag identification. This is different from [3], that introduces a LARS algorithm adapted to time series, where each variable can be represented by a matrix made of its lagged realisations : the algorithm manages to select iteratively blocks of lags corresponding to a single variable instead of single lags corresponding each to a variable. In the next section, a mathematical framework of this approach is given. The way linear state-space models are used to denoise variables is explained as well as the modeling of S&P 500 variations through the LagLasso procedure. In the last section, a backtesting of this method is provided : it is built on a sample period of 20 years, uses a gliding window of 5 years and a small number of macro-economic and financial indicators.

2 Presentation of the prediction method

More formally, we observe the predictors $x_t = (x_{i,t})_{1\leq i\leq d}$, where $x_{.,t} \in \mathbb{R}^d, \forall t$ and we want to forecast the real variable y_t at horizon h. The forecasting linear model is proposed :

$$\tilde{y}_{t+h} = \sum_{i=1}^{d} \beta_i \tilde{x}_{i,t-\sigma(i)}, \quad (*)$$

where $\tilde{y}_{t+h}$ and $\tilde{x}_{i,t}$ are the innovation residuals of the variable $x_{i,t}$ corresponding, for each i, to a given linear state space model, where $\sigma(i)$ is a lag corresponding to the i^{th} variable, and where $\beta = (\beta_1, \dots, \beta_d)$ is a real vector.

2.1 First step : Kalman filtering

We propose the following linear state space model :

$$\begin{cases} z_t = H_t\theta_t + v_t, & v_t \sim \mathcal{N}(0, V_t), \\ \theta_t = F_t\theta_{t-1} + w_t, & w_t \sim \mathcal{N}(0, W_t), \\ \theta_0 \sim \mathcal{N}(m_0, V_0), & \end{cases}$$

where $z_t \in \mathbb{R}^m$ stands for the observation vector, $\theta_t \in \mathbb{R}^p$ is a hidden random vector, H_t and F_t are real matrices of size respectively $m * p$ and $p * p$, and are to be specified. The only parameters of the model are the observation and evolution variances V (matrice of size $m * m$) and W (matrice of size $p * p$), that we estimate from available data using maximum likelihood.
The Kalman filter recursively estimates the internal state of the process θ_t given the sequence of noisy observations z_t. We denote by $\hat{\theta}_{t|t}$ the estimate of the state at time t given observations up to and including time t , and by $P_{t|t}$ the associated error covariance matrix. This can be summed up by the system of equations :

$$\begin{cases} \hat{\theta}_{t|t-1} = F_t\hat{\theta}_{t-1|t-1}, \quad (**) \\ P_{t|t-1} = F_tP_{t-1|t-1}F_t' + W_{k-1}, \\ r_t = z_t - H_t\hat{\theta}_{t|t-1}, \quad (***) \\ S_t = H_tP_{t|t-1}H_t' + V_t, \\ K_t = P_{t|t-1}H_t'S_t^{-1}, \\ \hat{\theta}_{t|t} = \hat{\theta}_{t|t-1} + K_tr_t, \\ P_{t|t} = (I - K_tH_t)P_{t|t-1}, \end{cases}$$

Equation (**) gives the predicted state at step t and equation (***) the innovation residual : this is the way we compute the quantities $\tilde{x}_t$ and $\tilde{y}_t$ stated in (*).
Finally, in our implementation, we use such a model for the response y_t and a single such model for all predictors $x_{i,t}$, for simplicity of use.

2.2 Second step : selecting variables and lags with LagLasso

Predicting y_{t+h} is achieved by selecting the most significant variables and lags, knowing that only one lag can be chosen per variable. We implemented a Lasso-type procedure : the LagLasso, which aims at building the vector β given in (*). From now on, we use the notation x_i for $\tilde{x_{i,t}}$ and y for $\tilde{y}_t$ and we use a double index for β to account for the variables and the lags. In addition, as for the Lasso, it is necessary to offer some criterion to choose a single step in this iterative process that determines a single vector $\hat{\beta}$: both C_p-type and cross-validation stopping criteria were considered.

3 Results and backtesting

In order to question the validity of this method and to explore possible refinements, several simple test methods are given. All of them are based on the same principle : considering the last 20 years of

LagLasso steps.

0. Choose lag^{max} and lag^{min} : $\sigma_i \in [lag^{min}, lag^{max}], \forall i$.
1. Standardization of the predictors $x_{i,\sigma}$ to have mean 0 and variance 1.
Initialisation : $r = y - \overline{y}$, $\beta_{i,\sigma} = 0, \forall i, \sigma$.
2. Find the predictor $x_{j,\sigma}$ most correlated with r.
3. Move $\beta_{j,\sigma}$ from 0 towards its least-squares coefficient $\langle x_{j,\sigma}, r\rangle$, until some other competitor $x_{k,\tau}$, $k \neq j$, has as much correlation with the current residual as does $x_{j,\sigma}$.
4. Move $(\beta_{j,\sigma}, \beta_{k,\tau})$ in the direction defined by their joint least squares coefficient of the current residual on $(x_{j,\sigma}, x_{k,\tau})$, until some other competitor $x_{l,\upsilon}$ has as much correlation with the current residual, i.e. : $< x_{l,\upsilon}, r >=< x_{k,\tau}, r >=< x_{j,\sigma}, r >$.
5a. If a non-zero coefficient hits zero, drop it from the active set, reinclude the variable and all its lags in the inactive set and recompute the current joint least squares direction.
5b. Eliminate all the lags corresponding to variable j from the inactive set. Continue until d variables are entered.

S&P 500 monthly variations, we use a gliding window containing a sufficient and constant number of points to make a prediction of the variation of the S&P 500 index over the next month. A number of successive predictions at horizon $h = 1\,\mathrm{month}$ are obtained and compared with those computed with other methods, linear state-space models and regression particularly.
Obviously, some explicative variables are needed and they have been chosen carefully : we checked that having too much correlated variables in the data basis is usually very counterproductive, which finally drastically limits the number of explicative variables. With the help of an economic expert, we chose PER, OIL, NAPM, INCOME and CORP PROFIT, that are all available on the website of the Federal Reserve Bank of St. Louis.
A first backtesting of this method consists in computing a recognition rate of upward and downward movements of the S&P 500 depending on the amplitude of the variation of the index. Results are provided for different maximal lags and for some other methods (cf. Table 1).
In addition, the following naive trading rule is proposed. Imagine a trader that decides to sell or to buy one unit of S&P 500 index every month. If the prediction of the model for next month is positive, the trader buys; if it is negative, he sells. At the end of the backtesting period, profit and loss accounts - computed with different maximal lags and following similar strategies derived from other methods - are compared (cf. Figure 1).

4 Conclusion

A first conclusion is that a multidimensional framework is usually more interesting than an unidimensional one. Furthermore, combining a filtering method with a selection method gives promising results : a simple state-space model combined with the Lasso outperforms all the other backtested methods. This has to be tempered by the delicate calibration of the model : if the database contains too much correlated variables, the results (recognition rate and profit and loss account) are clearly worse. And since macro-economic and financial variables usually strongly depend on each other, this limits the number of predictors in the database.
Unfortunately, taking lags into account does not improve significantly neither the recognition rate nor the profit and loss accounts, although it is a phenomenon highlighted by economists. We believe further improvements can be reached through a better indexing of data.

Table 1: recognition rate of S&P 500's upward and downward movements.

Amplitude of the variation	**Kalman and LagLasso,** $lag^{max} = 1$	**Kalman and LagLasso,** $lag^{max} = 6$	**Kalman and LagLasso,** $lag^{max} = 12$	**Level Model**	**Local Trend Model**	**Lasso**	**Regression**
0	61.6%	60%	62.2%	53.8%	56.1%	61.6%	61.6%
0.01	62.5%	59.8%	61.8%	55.2%	56.5%	61.8%	61.8%
0.02	64.4%	58.8%	60.7%	53.2%	58.8%	60.7%	62.6%
0.03	64.6%	62.1%	64.6%	53.6%	58.5%	60.7%	62.1%
0.04	66%	66%	66%	50.9%	64.1%	59.7%	62.2%
0.05	66.6%	71.7%	71.7%	51.2%	71.7%	58.4%	56.4%

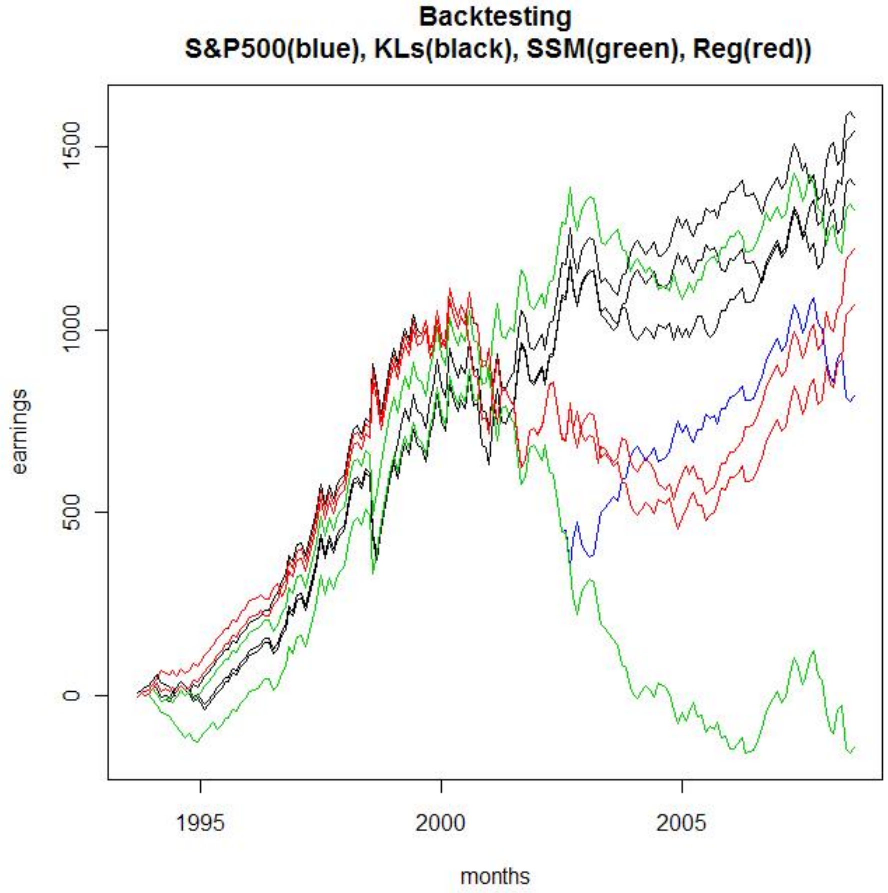

Figure 1: Backtesting

KL stands for Kalman/LagLasso-type methods($lag^{max} = 1, 6, 12$), SSM for linear State-Space Models (Level Model and Local Trend Model) and Reg for both Regression and Lasso.

5 References

[1] Efron, B.; Hastie, T.; Johnstone, I. and Tishirani, R. (2004), "Least Angle Regression," *The Annals of Statistics*, 32, 407-451.

[2] Hastie, T.; Taylor, J.; Tibshirani, R; and Walther, G. (2007), "Forward stagewise regression and the monotone lasso", *Electron. J. Statist.*, 1, 1-29.

[3] Croux, C; Gelper S. (2008), "Least Angle Regression for Time Series Forecasting with Many Predictors", 1-36 pp. Leuven: K.U.Leuven.

[4] Kalman, R. (1960), "A new approach to linear filtering and prediction problem", *Trans ASME Journal of Basic Engineering*, 35-45.

[5] Ruey, T., (2005), "Analysis of Financial Time Series", Wiley.

[6] Welch, G.; Bishop, G., "An Introduction to the Kalman Filter", University of North Carolina, Department of Computer Science, TR 95-041. 1995.

[7] Cappe, O., Moulines, E., Ryden, T. (2005), "Inference in Hidden Markov Models", *Springer Series in Statistics*, Springer-Verlag New York, Inc., Secaucus, NJ.

[8] Schleifer, A. (2000), "An Introduction to Behavioral Finance", Oxford University Press.