MODELING THE S&P 500 INDEX USING THE KALMAN FILTER AND THE LAGLASSO

Nicolas Mahler

*ENS Cachan & UniverSud
CMLA UMR CNRS 8536
*Telecom Paristech (TSI)
LTCI UMR Institut Telecom/CNRS 5141
*Strategic Risk Management
S.A., Paris

ABSTRACT

This article introduces a method to predict upward and downward monthly variations of the Standard & Poor's 500 (S&P 500) index by using a pool of macro-economic and financial explanatory 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 factors can efficiently represent the exogenous influence on S&P 500. The influence of each of these factors can change over time and it can be lagged. Additionnally, according to economists (cf. [1]), S&P 500 is sensitive to the movements of the factors around their own trend rather than to the variations themselves. Therefore, we need to filtrate the factors: a linear state-space model is first proposed for each of them and their innovation residuals are computed with the Kalman algorithm (cf.[2, 3, 4, 5]). 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 an L^2 regression fit under L^1 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 addressed by slightly modifying the Lasso, that is intimately connected to the LARS algorithm, which itself is an iterative procedure of variable selection generalizing the concept of bissector in a multidimensional framework(cf.[6] and [7]). It is important to note that this problem of lag identification is different from [8], 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. Our approach is thus 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. Actually, 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. The structure of the paper is the following. 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. Section 3 is devoted to the presentation of the database and to the appreciation of the relative importance and relative relevance of the variables. In the last section, several backtesting plans of this method are provided: all of them are built on a sample period of 20 years, use a gliding window of 5 years and a small number of macro-economic and financial indicators. Furthermore, comparisons to more classical methods with the same backtesting procedures are proposed.

2. PRESENTATION OF THE PREDICTION METHOD

We observe the random pair (x_t, y_t) over time t. We consider y_t as being the variable of interest and x_t is a random vector of dimension d whose components are explanatory variables $x_{i,t}, i = 1, \ldots d$. Considering $(x_t, y_t)_{t=1,\ldots,T}$, we aim at predicting $(y_t)_{t=T+1,\ldots,T+h}$, where the integer $h \geq 1$

978-1-4244-4948-4/09/$25.00 © 2009 IEEE

is the prediction horizon. In the paper, we propose to study a forecasting linear model of the following form:

$$y_{t+h} = \sum_{i=1}^{d} \beta_i x_{i,t-\sigma(i)}, \tag{1}$$

where $\sigma(i)$ which is a lag corresponding to the ith variable, $\beta = (\beta_1, \ldots, \beta_d)$ which is a real vector, are the unknown parameters to be estimated. Note that all variables considered are centered hence there is no constant term in the model.
The application we consider here is the prediction of upward and downward monthly variations of the S&P 500 index by using a pool of macro-economic and financial explanatory variables: y_t and $x_{i,t}$ are derived from these time series and stand for their innovation residuals defined in the sequel.

2.1. First step: Kalman filtering

We briefly recall the steps of Kalman filtering for linear state-space models (cf.[3]). We set the following notations: $z_t \in \mathbb{R}^m$ stands for the observation vector (in the application we consider, z_t will alternatively stand for y_t and $x_{i,t}$), $\theta_t \in \mathbb{R}^p$ is a hidden random vector, H_t and F_t which are real matrices of size respectively $m \times p$ and $p \times p$, are to be specified, and V_t and W_t are the observation and evolution covariance matrices of size respectively $m \times m$ and $p \times p$. The following linear state space model (2) is proposed:

$$\begin{cases} z_t = H_t\theta_t + v_t, & v_t \sim \mathcal{N}(0, V_t), \\ \theta_t = F_t\theta_{t-1} + w_t, & w_t \sim \mathcal{N}(0, W_t), \\ \theta_0 \sim \mathcal{N}(m_0, V_0), \end{cases} \tag{2}$$

In the following sections, we assume that $V_t = V$ and $W_t = W$, for any t. Then the observation and evolution covariance matrices V and W are the only parameters of the model. They are estimated from available data using maximum likelihood, thanks to the L-BFGS-B quasi-Newton optimization algorithm (cf. [9, 10]).
The Kalman filter recursively estimates the internal state of the process θ_t given the sequence of noisy observations z_t. We denote by $\hat{\theta}_{t|\tau}$ the estimate of the state at time t given observations up to and including time τ, and by $P_{t|\tau}$ the associated error covariance matrix. This can be summed up by the system of equations:

$$\begin{cases} \hat{\theta}_{t|t-1} = F_t\hat{\theta}_{t-1|t-1}, & (3) \\ P_{t|t-1} = F_tP_{t-1|t-1}F_t' + W_{k-1}, & \\ r_t = z_t - H_t\hat{\theta}_{t|t-1}, & (4) \\ S_t = H_tP_{t|t-1}H_t' + V_t, & \\ K_t = P_{t|t-1}H_t'S_t^{-1}, & \\ \hat{\theta}_{t|t} = \hat{\theta}_{t|t-1} + K_tr_t, & \\ P_{t|t} = (I - K_tH_t)P_{t|t-1}, & \end{cases}$$

Equation (3) gives the predicted state at step t and equation (4) the innovation residual : this is the way the quantities x_t and y_t stated in (1) are computed.
In our case of multivariate time series, one could wonder whether to use in practice a different filter for each individual time series or the same filter. In our implementation, we used one model for the response y_t and the same model for all covariates $x_{i,t}$ for simplicity of use.

2.2. Second step: selecting variables and lags with Lag-Lasso

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 (1). This procedure is iterative and requires to manage a set of covariates, that we call the active set. From now on, we use the notation x_i for $x_{i,t}$ and y for y_t and we use a double index for β to account for the variables and the lags. The steps of the LagLasso are the following:

1. *Choose* $\text{lag}^{\max}$ *and* $\text{lag}^{\min}$*:* $\sigma_i \in [\text{lag}^{\min}, lag^{max}], \forall i$.
2. *Standardize the covariates* $x_{i,\sigma}$ *to have mean* 0 *and variance* 1*. Initialisation:* $r = y - \overline{y}$*,* $\beta_{i,\sigma} = 0, \forall i, \sigma$.
3. *Find the covariates* $x_{j,\sigma}$ *most correlated with* r.
4. *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* $x_{j,\sigma}$ *has.*
5. *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.:* $< x_{l,\upsilon}, r >=< x_{k,\tau}, r >=< x_{j,\sigma}, r >$.
6. *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.*
7. *Eliminate all the lags corresponding to variable* j *from the inactive set. Continue until* d *variables are entered.*

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. PRESENTATION OF THE DATABASE AND SELECTION OF VARIABLES.

The explanatory variables we needed have to be chosen carefully. We checked that having too many correlated variables in the data base is usually very counterproductive, which

drastically limits the number of explanatory variables. Thus, starting from an important pool of financial and macro - economic (monthly sampled) variables, a first hint concerning the variables to eliminate is given by the correlation between filtered variables. With the help of an economic expert, we then restricted ourselves to the following US data:

- the S&P 500 (**S&P500**).
  The S&P 500 index is a value weighted index composed of the prices of 500 common stocks, all of them nearly being stocks of 500 most important companies trading in the American stock markets (a stock of a company is a share of the ownership of that company represented in a negociable and fungible financial instrument).
- the S&P 500 Price-to-Earnings Ratio (**PER**).
  The P/E ratio of a stock is a measure of the price paid for a share relative to the annual net income or profit earned by the firm per share. In the case of a market index such as the S&P 500, a weighted average of all stock constiutents is computed, by summing each stock's underlying market cap (price multiplied by number of shares in issue) to give the total value in terms of market capitalization for the whole market index.
- the Spot Oil Price, West Texas Intermediate, in dollars per barrel (**OIL**).
  West Texas Intermediate is a type of crude oil used as a benchmark in oil pricing and the underlying commodity of New York Mercantile Exchange's oil futures contracts.
- the NAPM, ISM Manufacturing, Purchasing Managers' Composite Index (**PMI**).
  Released by the Institute for Supply Management every month this economic indicator is a composite index that is based on five major indicators including: new orders, inventory levels, production, supplier deliveries, and the employment environment. It is the best indicator of factory production and it is popular for detecting inflationary pressure as well as manufacturing economic activity, both of which investors pay close attention to.
- the Disposable Personal Income (**INCOME**), i.e. gross income minus income tax on that income.
  When applying for a loan (mortgage, consumer loan), lenders may take into consideration a high-income applicant's income in order to assess the loan repayment capacity of the applicant.
- the Corporate Profits After Tax (**CORP PROFIT**).
  Corporate Profits are the income of organizations treated as corporations in the national income and product accounts. They are the lifeblood of investment spending: when profits are strong, then companies will be able to increase their capital spending. The also reflect the efficiency of an organization: for exeample, when a company's profits are relatively strong, even during an economic downturn, it usually means that the organization is well-managed.
- the US Population (**POPULATION**).
- the US Unemployment rate (**UNEMPLOYED**).

All of them are US data, available on the website of the Federal Reserve Bank of St. Louis (for further information, cf. [11, 12, 13, 14, 15]).

As stated in the introduction, all the performed backtesting plans use a gliding window of 5 years on a period of 20 years: the figure 1 shows the occurences of each variable over that period for a specific backtesting explained in the next part. This gives a hint about the relative importance of the variables but not necessarily about their relative relevance. To check that, Table 1 is produced next page. Finally, one can also build a histogram giving the occurences of the lags, as shown in Figure2.

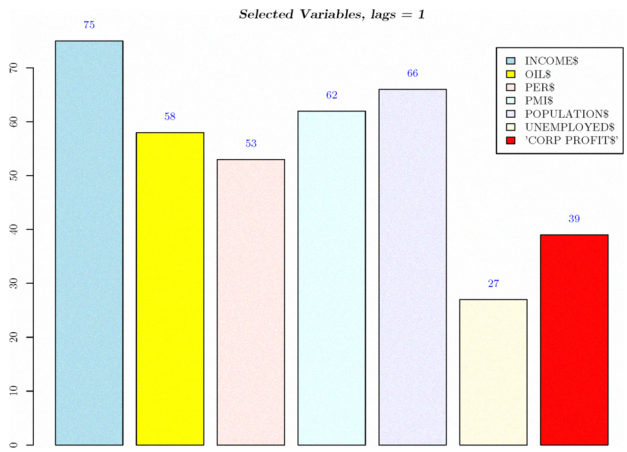

Fig. 1. Selected Variables Histogram for **Kalman LagLasso,** $\text{lag}^{\max} = 1$.

Table 1. Right and wrong predictions per variable for **Kalman LagLasso,** $\text{lag}^{\max} = 1$.

Explanatory variable	**number of selections**	**number of right predictions**	**number of wrong predictions**
INCOME	75	38	37
OIL	58	28	30
PER	53	21	32
PMI	62	35	27
POPULATION	66	33	33
UNEMPLOYED	27	11	16
CORP PROFIT	39	28	11

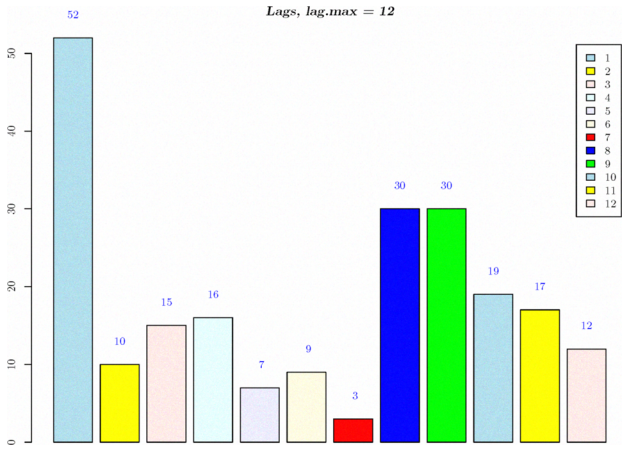

Fig. 2. Selected Lags Histogram for **Kalman LagLasso,** $\text{lag}^{\max} = 12$.

4. RESULTS AND BACKTESTING

In order to question the validity of Kalman LagLasso procedures 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 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 $h = 1\,\text{month}$ are obtained and compared with those computed with the following methods:

- **State-Space Model 1**, also called local level model, or Arima(0,1,1): in the linear state space model (2), let $H_t = F_t = 1$. No explanatory variables are used. SSM1 is represented in green on Figure 3.
- **State-Space Model 2**, also called local linear trend model, or Arima(0,2,2): in the linear state space model (2), let $F_t = 1$ and $H_t = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$. No explanatory variables are used. SSM2 is also in green on Figure 3.
- **Kalman LagLasso-based models** for $\text{lag}^{\max} = 1$, $\text{lag}^{\max} = 6$, $\text{lag}^{\max} = 12$: the S&P 500 only is filtered using SSM2, while all the explanatory variables are filtered using SSM1 1. Kalman LagLasso-based models are represented in black on Figure 3.
- **Regression-based** and **Lasso-based models**: the S&P 500 only is filtered using SSM2 while the explanatory variables are not; selection of variables is performed through Regression or Lasso. Both models are represented in red on Figure 3.

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 2): LagLasso procedures clearly outperform the other ones. Note that we also use that backtesting procedure for method *Kalman LagLasso* with $\text{lag}^{\max} = 1$ to compute the overall histogram 1 and to produce Table 2, that gives per variable the number of right predictions and the number of wrong predictions. Table 1 shows that despite the good results obtained by LagLasso procedures, the recognition rate of nearly every variable is quite poor. In fact, a careful examination of the selected variables over time shows that the right predictions given by each variable tend to be of greater amplitude than the wrong ones. There is still an effort to be made concerning the sparsity of the selection. Furthermore, 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 3): gains obtained with LagLasso procedures are all much higher than gains obtained with other methods. Both backtesting plans are moreover completed by Figure 4 giving the kernel estimated density of the conditional variables $y_t|(sign(\tilde{y}_t) = 1)$ and $y_t|(sign(\tilde{y}_t) = -1)$, where y_t stands for the observation derived from S&P500 and $\tilde{y}_t$ is the prediction delivered by Kalman LagLasso procedure with $\text{lag}^{\max} = 12$. It shows that the Kalman LagLasso procedure is particularly efficient to detect downward trends of great amplitude, but performs poorly when confronted to upward trends of great amplitude.

As a conclusion, although our LagLasso procedure gives very promising results and outperform more classic methods for both proposed backtesting plans, difficulties remain concerning:

- the sparsity of selection: as mentionned above, variables are sometimes selected when not necessary, having a weak influence in the final prediction, and inducing thus a poor recognition rate of downward and upward trends per variable.
- above all, the phenomenon of lags does exist according to economists but is hardly highlighted here, as the results between Kalman LagLasso procedures for different maximal lags are very close.
- the asymmetry of the efficiency of the predictions: although downward trends of great amplitude are well forecasted, important upward trends are not.

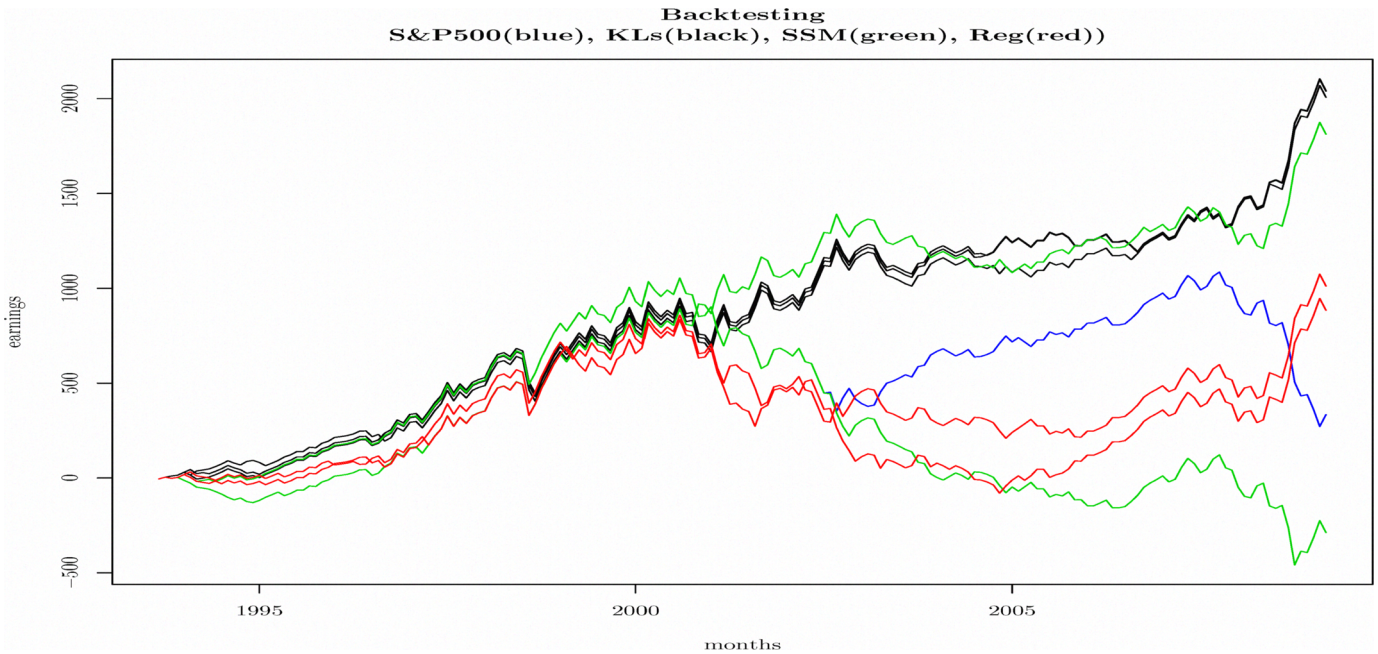

Fig. 3. Backtesting.

Table 2. Recognition rate of S&P 500's upward and downward movements.

Amplitude of the variation	**Kalman LagLasso,** $\text{lag}^{\max} = 1$	**Kalman LagLasso,** $\text{lag}^{\max} = 6$	**Kalman LagLasso,** $\text{lag}^{\max} = 12$	**Level Model**	**Local Trend Model**	**Lasso**	**Regression**
0	59.8%	62%	62.5%	53.4%	56.6%	56.6%	57.2%
0.01	60.7%	63.2%	63.2%	55%	57.5%	58.2%	57.5%
0.02	59.2%	62.8%	61.9%	53%	60.1%	58.4%	58.4%
0.03	63.6%	67%	65.9%	53.4%	60.2%	56.8%	56.8%
0.04	69.4%	67.8%	67.8%	50.8%	66.1%	59.3%	57.6%
0.05	75.5%	73.3%	73.3%	51.1%	73.3%	62.2%	55.5%

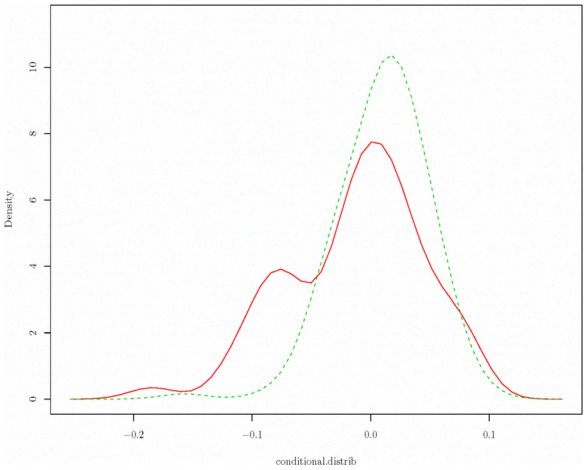

Fig. 4. SP500 observations knowing predictions.

5. CONCLUSION

A first conclusion is that a multidimensional framework is usually more interesting than an one-dimensional one. Furthermore, combining a filtering method with a selection method gives promising results: a simple state-space model combined with the LagLasso outperforms all the other backtested methods, for both proposed backtesting plans. This has to be tempered by the delicate calibration of the model: if the database contains too many correlated variables, the results (recognition rate and profit and loss account) are clearly worse. Since macro-economic and financial variables usually strongly depend on each other, this limits the number of factors in the database. Moreover, taking lags into account does unfortunately 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 the data.

6. REFERENCES

[1] A. Schleifer, *An Introduction to Behavioral Finance*, Oxford University Press, 2000.

[2] R. Kalman, "A new approach to linear filtering and prediction problem," *Trans ASME Journal of Basic Engineering*, pp. 35–45, 1960.

[3] T. Ruey, *Analysis of Financial Time Series*, Wiley, 2005.

[4] G. Welch and G. Bishop, "An introduction to the kalman filter," 2001.

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

[6] I. Johnstone B. Efron, T. Hastie and R. Tishirani, "Least angle regression," *The Annals of Statistics*, vol. 32, no. 2, pp. 407–451, 2004.

[7] R. Tishirani B. Efron, J. Taylor and G. Walther, "Forward stagewise regression and the monotone lasso," *Electron. J. Statist.*, vol. 1, pp. 1–29, 2007.

[8] S. Gelper C. Croux, "Least angle regression for time series forecasting with many predictors," *K.U.Leuven*, pp. 1–36, 2008.

[9] D. C. Liu and J. Nocedal, "On the limited memory bfgs method for large scale optimization," *Math. Program.*, vol. 45, no. 3, pp. 503–528, 1989.

[10] J. Nocedal R. H. Byrd, P. Lu and C. Zhu, "A limited memory algorithm for bound constrained optimization," *SIAM J. Sci. Comput.*, vol. 16, no. 5, pp. 1190–1208, 1995.

[11] Jeremy J. Siegel, *Stocks for the Long Run*, McGraw-Hill, 2007.

[12] Institute for Supply Management, "March 2009 manufacturing ism report on business," http://www.ism.ws/.

[13] Bureau of Economic Analysis, "Regional definitions," http://www.bea.gov/regional/definitions/.

[14] Bureau of Economic Analysis, "National economic accounts," http://www.bea.gov/national/index.htm.

[15] Federal Reserve Bank Of St. Louis, "Economic research," http://research.stlouisfed.org/.