Non-linear Supervised High Frequency Trading Strategies with Applications in US Equity Markets

Nan Zhou, Wen Cheng, Ph.D.

Associate, Quantitative Research, J.P. Morgan
nan.zhou@jpmorgan.com

The 4th Annual Modeling High Frequency Data in Finance Conference
July 2012
Outline

1. Overview
2. Data and Descriptive Analysis
3. (SKIPPED) Linear Model: Variable Selection by LASSO
4. MARS: Multivariate Adaptive Regression Splines
5. GBM: Generalized Boosted Models
6. More Results based on SPY 2011 December Level-1 Data
7. Discussions
### Disclosures

- This talk is for informational purposes only. Any comments or statements made herein do not necessarily reflect those of JPMorgan Chase & Co., its subsidiaries and affiliates.

- Illustrations and discussions of **predictive models** based on High Frequency Financial Data

- Nonlinear \( (\beta x \rightarrow \sum \beta_i f(x, \theta_i)) \); Supervised (fixed holding period)

- Simple trading strategies based on high frequency predictions

### References (Supervised vs. Unsupervised):


# Level-1 S&P500 Emini Future data

<table>
<thead>
<tr>
<th>Series</th>
<th>Date</th>
<th>Type</th>
<th>Price</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7/18/2012 9:00:00 AM</td>
<td>ASK</td>
<td>1353.50</td>
<td>59</td>
</tr>
<tr>
<td>2</td>
<td>7/18/2012 9:00:00 AM</td>
<td>BID</td>
<td>1353.25</td>
<td>274</td>
</tr>
<tr>
<td>3</td>
<td>7/18/2012 9:00:00 AM</td>
<td>BID</td>
<td>1353.25</td>
<td>275</td>
</tr>
<tr>
<td>4</td>
<td>7/18/2012 9:00:00 AM</td>
<td>BID</td>
<td>1353.25</td>
<td>277</td>
</tr>
<tr>
<td>5</td>
<td>7/18/2012 9:00:00 AM</td>
<td>ASK</td>
<td>1353.50</td>
<td>58</td>
</tr>
<tr>
<td>6</td>
<td>7/18/2012 9:00:00 AM</td>
<td>ASK</td>
<td>1353.50</td>
<td>54</td>
</tr>
<tr>
<td>.......</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>996116</td>
<td>7/18/2012 1:59:59 PM</td>
<td>ASK</td>
<td>1367.50</td>
<td>485</td>
</tr>
<tr>
<td>996117</td>
<td>7/18/2012 1:59:59 PM</td>
<td>BID</td>
<td>1367.25</td>
<td>477</td>
</tr>
<tr>
<td>996118</td>
<td>7/18/2012 1:59:59 PM</td>
<td>BID</td>
<td>1367.25</td>
<td>478</td>
</tr>
<tr>
<td>996119</td>
<td>7/18/2012 2:00:00 PM</td>
<td>BID</td>
<td>1367.25</td>
<td>477</td>
</tr>
<tr>
<td>996120</td>
<td>7/18/2012 2:00:00 PM</td>
<td>BID</td>
<td>1367.25</td>
<td>476</td>
</tr>
<tr>
<td>996121</td>
<td>7/18/2012 2:00:00 PM</td>
<td>ASK</td>
<td>1367.50</td>
<td>483</td>
</tr>
<tr>
<td>.......</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Target:

Predicting the price movement in the next $H$ seconds using historical information.

- Supervised Learning: Regression, Neural Network, Boosting, SVM, etc;
- Unsupervised Learning: Clustering, Graphical Model (Network), Reinforcement Learning, etc.
Inputs & Outputs:

\[
r_{t-l}^h = S_{t-l} - S_{t-l-h}, \quad S_t = \frac{\text{bid}_t + \text{ask}_t}{2}
\]

Outputs (forward absolute returns):

- \( fr_t^H = S_{t+H} - S_t \) (e.g. \( H = 300 \) s)

Inputs (absolute returns):

- Existing variables:
  - Group1: \( \{r_t^1, r_t^{H/2}, r_t^H\} \)
  - Group2: \( \{r_t^{H/m}\}, \quad m = 1, 2, 5, 10, 20, 50, 100, 300 \)
  - Group3: \( \{r_t^{H/m}\}, \quad m = 1, \ldots, H \)
- Technical indicators: RSI, MACD, Bollinger Band, etc;
- Potential variables: size, volatility, economic data, other stock/index prices, expert’s analysis, etc.
A Simple High Frequency Trading Strategy:

Predictions at time $t$: $\hat{fr}_t^H = S_{t+H} - S_t = f \left( \{ r_{t-l}^h \} \right)$

**Entry signals:**
- Open a market **buy** order at time $t$, if $\hat{fr}_t^H \geq$ cost;
- Open a market **sell** order at time $t$, if $\hat{fr}_t^H \leq -$cost;
- No action, otherwise.

**Exit signals:**
- **Liquidate** the position (opened at time $t$) using market order at time $t+H$. 
Outline

1. Overview

2. Data and Descriptive Analysis

3. (SKIPPED) Linear Model: Variable Selection by LASSO

4. MARS: Multivariate Adaptive Regression Splines

5. GBM: Generalized Boosted Models

6. More Results based on SPY 2011 December Level-1 Data

7. Discussions
Time Series:

- Middle Price of ER 2007-07-02
- Autocorrelation
- Return (in 1 sec) of Mid Price
- Return (in 300 secs) of Mid Price
Distributions of Different Returns:
Scatter Plots: $Y_t = S_{t+300} - S_t$ v.s. $r^h_t = S_t - S_{t-h}$
Outline

1. Overview

2. Data and Descriptive Analysis

3. (SKIPPED) Linear Model: Variable Selection by LASSO

4. MARS: Multivariate Adaptive Regression Splines

5. GBM: Generalized Boosted Models

6. More Results based on SPY 2011 December Level-1 Data

7. Discussions
Generalized Linear Model

Model Components

**Response/Dependent Variable:** $Y$

**Predictors/Independent Variables:** $X_1, \ldots, X_p$

**Observations:** $y_{n \times 1} = [y_1, \ldots, y_n]'$, $x_{n \times p} = [x_1, \ldots, x_p]$

Model Setup

**Model:** $Y|X \sim F_Y$, $E(Y) = \mu = g^{-1}(X\beta)$, $Var(Y) = V(g^{-1}(X\beta))$

where, $\beta_{p \times 1}$ is unknown parameters; $g$ is the link function.

Examples

- **Linear Regression:** $Y|X \sim N(X\beta, \sigma^2)$, or  
  
  $y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i$, where $\epsilon_i \sim i.i.d \sim N(0, \sigma^2)$

- **Logistic Regression:** $Y|X \sim \text{Bernoulli}(p = g^{-1}(X\beta))$, $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$, or  
  
  $P(y_i = 1) = \frac{\exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}$
Generalized Linear Model

Parameter Estimations
- Maximum Likelihood, Maximum Quasi-Likelihood, Minimum Loss Function
- Least Square, Iteratively Reweighted Least Squares
- Bayesian Methods

Overfitting
Overfitting:

### Traditional Methods
- Forward, Backward, Stepwise Selection
- Best Subset Selection: $R^2$, MSE, AIC, SBC
- Principal Components Regression
- Ridge Regression: Shrinkage Estimations

### Penalized Methods
- LASSO, SCAD, Dantzig selector
- SIS, ISIS
Ridge Regression-L2 Penalized Methods

\[
y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i, \epsilon_i \text{ i.i.d } \sim N(0, \sigma^2)
\]

\[L(X, \beta)\]-Loss Function

**Least Squares Regression**

\[
\hat{\beta}_{ls}^\sim = \arg\min_{\beta} \{L(X, \beta)\}
\]

\[
= \arg\min_{\beta} \{\sum_{i=1}^{p} (y_i - \hat{y}_i)^2\}
\]

\[
\hat{\beta}_{ls}^\sim = (X'X)^{-1}Y
\]

**Ridge Regression**

\[
\hat{\beta}_{ridge}^\sim = \arg\min_{\beta} \{L(X, \beta)\}
\]

\[
= \arg\min_{\beta} \{\sum_{i=1}^{p} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2\}
\]

\[
\hat{\beta}_{ridge}^\sim = (X'X + \lambda I)^{-1}Y
\]
Ridge Regression-L2 Penalized Methods

\[ y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i, \epsilon_i \text{ i.i.d } \sim N(0, \sigma^2) \]

\[ L(X, \beta) \text{-Loss Function} \]

**Least Squares Regression**

\[ \hat{\beta}_{ls} = \arg\min_{\beta} \{ L(X, \beta) \} \]

\[ = \arg\min_{\beta} \{ \sum_{i=1}^{p} (y_i - \hat{y}_i)^2 \} \]

\[ \hat{\beta}_{ls} = (X'X)^{-1}Y \]

**Ridge Regression**

\[ \hat{\beta}_{ridge} = \arg\min_{\beta} \{ L(X, \beta) \} \]

\[ = \arg\min_{\beta} \{ \sum_{i=1}^{p} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \} \]

\[ \hat{\beta}_{ridge} = (X'X + \lambda I)^{-1}Y \]
Ridge Regression-L2 Penalized Methods

\[ \hat{\beta}_{ridge} = \arg\min_{\beta} \{ \| \tilde{Y} - X \tilde{\beta} \|^2 + \lambda \| \tilde{\beta} \|_2 \}, \quad \text{where} \quad \| \tilde{\beta} \|_2 = \sum_{i=1}^{p} |\beta_i|^2 \]

\[ \Leftrightarrow \]

\[ \hat{\beta}_{ridge} = \arg\min_{\beta} \{ \| \tilde{Y} - X \tilde{\beta} \|^2 \} \]

\[ \text{s.t.} \quad \| \tilde{\beta} \|_2 \leq t \]

There is a one-to-one correspondence between \( \lambda \) and \( t \)
Ridge Regression-L2 Penalized Methods
LASSO-L1 Penalized Model

**LASSO (Least Absolute Shrinkage and Selection Operator):**

\[ y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i, \; \epsilon_i \; \text{i.i.d} \sim N(0, \sigma^2) \]

\[ \hat{\beta}_{lasso} = \arg\min_{\beta} \{ \|Y - X\beta\|^2 + \lambda \|\beta\|_1 \}, \text{ where } \|\beta\|_1 = \sum_{j=1}^{p} |\beta_j| \]

\[ \iff \]

\[ \hat{\beta}_{lasso} = \arg\min_{\beta} \{ \|Y - X\beta\|^2 \}
\quad \text{s.t. } \|\beta\|_1 \leq t \]
LASSO-L1 Penalized Model
Comparison of L1 and L2 Penalized Model

**L2 Penalized Estimation**

\[
\hat{\beta}^{\text{ridge}} = \arg\min_{\beta} \{ \| Y - X \hat{\beta} \|^2 \}
\]

s.t. 
\[\| \hat{\beta} \|_2 \leq t\]

**L1 Penalized Estimation**

\[
\hat{\beta}^{\text{lasso}} = \arg\min_{\beta} \{ \| Y - X \hat{\beta} \|^2 \}
\]

s.t. 
\[\| \hat{\beta} \|_1 \leq t\]
LASSO

References:

- **The original paper:**

- **The least angle regression (LAR) algorithm for solving the Lasso:**

- **Details and comparisons:**
  Hastie, T., Tibshirani, R. and Jerome, F. *The elements of statistical learning*, second edition, Springer

Computations:

- LASSO in R: glmnet, lasso2, lars
- Relaxed LASSO in R: relaxo
LASSO

References:

- **The original paper:**

- **The least angle regression (LAR) algorithm for solving the Lasso:**

- **Details and comparisons:**
  Hastie, T., Tibshirani, R. and Jerome, F. *The elements of statistical learning*, second edition, Springer

Computations:

- LASSO in R: **glmnet**, lasso2, lars
- Relaxed LASSO in R: relaxo
Outline

1. Overview
2. Data and Descriptive Analysis
3. (SKIPPED) Linear Model: Variable Selection by LASSO
4. MARS: Multivariate Adaptive Regression Splines
5. GBM: Generalized Boosted Models
6. More Results based on SPY 2011 December Level-1 Data
7. Discussions
Ideas:

- **Linear Regression:** \[ Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \epsilon_i \]

- **MARS:** Instead of \( X_j \), MARS uses a collection of new predictors in the form of piecewise linear basis functions:

\[ \{(X_j - t)_+, (t - X_j)_+\}, \ j = 1, \cdots, p, \ t \in \{x_{1j}, \cdots, x_{Nj}\} \]

Algorithm:

1. **Stagewise forward selection:**
   Keeping coefficients the same for variable existed in the current model, select a new basis function pair that produces the largest decrease in training error. Repeat the whole process until some termination condition is met:
   - reached the maximum number (pre-determined) of predictors;
   - adding any new predictor changes RSquare by less than 0.001;
   - reached a very high RSquare value.

2. **Backward pruning:**
   find the subset which gives the lowest Cross Validation error, or GCV.
Ideas:

- **Linear Regression:** \( Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \epsilon_i \)

- **MARS:** Instead of \( X_j \), MARS uses a collection of new predictors in the form of piecewise linear basis functions:
  \[ \{(X_j - t)_+, (t - X_j)_+\}, \ j = 1, \cdots, p, \ t \in \{x_{1j}, \cdots, x_{Nj}\} \]

Algorithm:

1. **Stagewise forward selection:**
   Keeping coefficients the same for variable existed in the current model, select a new basis function pair that produces the largest decrease in training error. Repeat the whole process until some termination condition is met:
   - reached the maximum number (pre-determined) of predictors;
   - adding any new predictor changes RSquare by less than 0.001;
   - reached a very high RSquare value.

2. **Backward pruning:**
   find the subset which gives the lowest Cross Validation error, or GCV.
Cross Validation:
Model Fitting: \( Y = \sum_j \beta_j f(X_j) \)
Results:

Trading Records of ER: MARS, predictors from group2
Holding period = 300 seconds

Training Period

Testing Period

Price (in cent)

Timestamp (by 1 seconds)
Results:
Results:

Trading Records of ER: MARS, predictors from group2
Holding period = 300 seconds

Training Period

Testing Period

Price (in cent)

Timestamp (by 1 seconds)
Results:
Results:
Outline

1. Overview
2. Data and Descriptive Analysis
3. (SKIPPED) Linear Model: Variable Selection by LASSO
4. MARS: Multivariate Adaptive Regression Splines
5. GBM: Generalized Boosted Models
6. More Results based on SPY 2011 December Level-1 Data
7. Discussions
History of Boosting (Trevor Hastie, January 2003):

- L.G. Valiant (1984)
- R.E Schapire (1990)
- Y. Freund (1995)
- Freund & Schapire (1996)
- R. Quinlan (1996)
- Breiman (1996, 1997)

Genesis of PAC Learning Model
Concept of Boosting appears
AdaBoost is born
Experiments with AdaBoost
Attempts to explain why AdaBoost works
Improvements

- Schapire, Y. Singer (1998)
- Friedman, Hastie, Tibshirani (1998)
- L. Breiman (1996) Bagging
Adaboost:

1. Initialize the observation weights: $w_i = 1/N, \ i = 1, \cdot \cdot \cdot, N$
2. For $m = 1$ to $M$,
   - Fit a classifier $G_m(x, \theta)$ to the training data using weights $w_i$
     - Stump, CART, ADT tree, etc;
   - Compute the miss-prediction error by:
     $$err_m = \sum w_i \cdot I(y_i \neq G_m(x_i), \theta)$$
   - Compute $\alpha_m = \log \frac{1-err_m}{err_m}$
   - Update the weights: $w_i \leftarrow w_i \cdot \exp(\alpha_m \cdot I(y_i \neq G_m(x_i, \theta)))$
3. Final Output: $G(x) = \text{sign} [\sum \alpha_m \cdot G_m(x, \theta)]$
Equivalence: Additive Model with Exponential Loss (Forward Stagewise Algorithm):

\[ G(x) = \sum_{m=1}^{M} \beta_m \cdot f(x, \theta_m) \]

**Forward Stagewise Algorithm**

1. Initialize \( G(x) = 0 \);
2. For \( m = 1 \) to \( M \),
   - Calibrate the model:
     \[ (\beta_m, \theta_m) = \arg\min_{\beta,\theta} \sum_{i=1}^{N} L(y_i, G(x_i) + \beta \cdot f(x, \theta)) \]
   - Update: \( G(x) \leftarrow G(x) + \beta_m \cdot f(x, \theta_m) \)

**Exponential Loss:** \( L(y, f(x)) = \exp(-y \cdot f(x)) \)
**Exponential Loss (Hastie, 2008):**

\[
L(y, f(x)) = \exp(-y \cdot f(x))
\]
Gradient Boosting:

Forward Stagewise Algorithm

1. Initialize $G(x) = \text{argmin}_a \sum_{i=1}^{N} L(y_i, a)$;
2. For $m = 1$ to $M$,
   - Calculate $r_i = -\left[ \frac{\partial L(y,f)}{\partial f} \right] |_{f=G(x_i)}$
   - Fit a regression tree based on $\{r_i\}$ giving terminal regions $\{R_1, \cdots, R_K\}$
   - for $k = 1$ to $K$, compute
     $$ a_k = \text{argmin}_a \sum_{x_i \in R_k} L(y_i, f(x_i) + a) $$
   - Update: $G(x) \leftarrow G(x) + \sum_k a_k I(x \in R_k)$
Model Fitting:
Results:
Outline

1 Overview

2 Data and Descriptive Analysis

3 (SKIPPED) Linear Model: Variable Selection by LASSO

4 MARS: Multivariate Adaptive Regression Splines

5 GBM: Generalized Boosted Models

6 More Results based on SPY 2011 December Level-1 Data

7 Discussions
Results based on SPY 2012 December Level-1 Data:
Outline

1. Overview
2. Data and Descriptive Analysis
3. (SKIPPED) Linear Model: Variable Selection by LASSO
4. MARS: Multivariate Adaptive Regression Splines
5. GBM: Generalized Boosted Models
6. More Results based on SPY 2011 December Level-1 Data
7. Discussions
Discussions:

- Generally speaking, our principles to build a successful quantitative trading strategy include:
  - Modeling the market sense and experience quantitatively (ex: which information is useful and could be coded into the predictors, what is the rationale for a strong correlation, etc.);
  - Keeping the model as simple as possible and generalizing the model intuitively instead of overfitting by the 'magic' black-box.
- Dynamic/online Trading Strategy; Window sizes for training and testing
- Choice of holding period; Expert System; Asset Allocations
- Homogeneity: different intra-day time periods
- Additional predictors; different definitions of price return
- Risk Management
Questions?