Evidence Optimization for Consequently Generated Models

Vadim Strijov\textsuperscript{a,}\textsuperscript{*}, Katya Krymova\textsuperscript{a}, Gerhard Wilhelm Weber\textsuperscript{b}

\textsuperscript{a}Computing Center of the Russian Academy of Sciences, Moscow, Russia
\textsuperscript{b}Institute of Applied Mathematics, Middle East Technical University, Ankara, Turkey

Abstract
To construct an adequate regression model one has to fulfill the set of measured features with their generated derivatives. Often the number of these features exceeds the number of the samples in the data set. After a feature generation process the problem of feature selection from a set of highly correlated features arises. The proposed algorithm uses an evidence maximization procedure to select a model as a subset of generated features. During the selection process it rejects multicollinear features. A problem of European option volatility modelling illustrates the algorithm. Its performance is compared with performance of similar well-known algorithms.

Keywords: European option volatility, feature generation, model evidence, model selection, multicollinearity

1. Introduction

We investigate a linear regression model construction process. To construct an adequate model that fits data well one has to generate additional features [1, 2] and select the most informative ones [3]. The term “the most informative” means that this feature brings optimal value to some given quality criterion. This criterion depends on statistical hypothesis or could be assigned externally.

In this paper, the model evidence is used as the model quality criterion. It is based on the following hypothesis: the normal distribution of target vector in the linear model implies the normal distribution of model parameters. The model evidence estimation is a part of the coherent Bayesian inference, described in [4–6]. It consist of two levels of inference. The first level compares the competitive models; the second level estimates the most probable parameters of a model.

To detect multicollinear features one can check correlations between features [7], examin the variance inflation factor [8] or use the Belsley method [9]. The last one combines singular value decomposition technique [10–12] and estimation of the model parameters variance [13].

The comparative study investigates three groups of feature selection algorithms: exhaustive search algorithms, regularization algorithms and “add” algorithms. The first group includes some external criterion to maximize during the model selection process. It includes exhaustive search [14], genetic algorithm [15] and group methods of data handling [14].

The second group applies additional restriction to the model parameters or to model quality criterion. It includes ridge regression [7], and Tibshirani’s Lasso algorithm [16].

The last group includes stagewise regression [7], forward orthogonal search [17] and least angle regression, LARS [18]. These algorithms use their proper criterions of feature addition to control the sequence of selected features. The group also includes very popular “add-del” stepwise regression algorithm [19].

Illustrate the main defect of the “add-del” algorithms with the following example. Let the design matrix $X$ consists of three features. The first one, $x_1$ is highly correlated with the target vector $y$; however the target vector could be represented as a linear combination of the vectors $x_2$, $x_3$. For example,

$$X = [x_1, x_2, x_3] = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \quad \text{and} \quad y = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix},$$

see Figure 1.

![Figure 1: Illustration of consequent feature addition](image)

“Add”-algorithms like LARS and FOS select the feature $x_1$ on the first step, since this feature correlates to the target vector well, and then select the rest of the features. The quality criterion $S \neq 0$ for the complexity $c_f = 2$. However in this case there exists a way to bring $S = 0$ for the same complexity.

The following algorithm is proposed to circumvent this difficulty. There are two main requirements to consequently gener-
ated models:

1) the evidence must be maximal,
2) the multicollinearity must be minimal.

In this paper we propose a new model selection algorithm. The proposed algorithms includes multicollinearity analysis and obtains well-conditioned feature sets. The algorithm is based on consequent addition and deletion of the features and constructs linear regression model of maximal evidence. During the “add” part it consequently increases number of features in the model until the model will be overtrained (see Figure 2 for illustration).

Comparison with the LARS algorithm shows that multicollinearity detection allows to decrease conditional number of the design matrix without significant decay of the model quality.

The paper contain four parts. The second part states a problem of classical model selection. The third part describes a way to generate derivative features, that could significantly improve the model quality. The fourth part is devoted to the model evidence as an alternative criterion of model quality. The fifth part shortly describes the Belsley method of multicollinearity detection and name it the active set. Assume the model class is put in one-to-one correspondence with the superposition \( \mathcal{F}(\xi_0) \). Consider the Cartesian product \( G \times \Xi \). Each element \( (g_\varepsilon, \xi_0) \) of \( G \times \Xi \) is put in one-to-one correspondence with the superposition \( g_\varepsilon(\xi_0) \). Denote by \( a_\varepsilon \) the superposition \( g_\varepsilon(\xi_0) \), where \( \varepsilon = (v - 1)U + u \).

A polynomial in variables \( a_\varepsilon \) is the linear model (1) from the set \( \mathcal{F} \). Represent this polynomial in the form

\[
y = w_0 + \sum_{i=1}^{UV} w_i a_i + \sum_{i=1}^{UV} \sum_{\xi=1}^{UV} w_{ij} a_i a_{\xi} + \ldots + \sum_{i=1}^{UV} \sum_{\xi=1}^{UV} \sum_{\zeta=1}^{UV} \sum_{\eta=1}^{UV} \cdots w_{ij\xi\zeta\eta} a_i a_{\xi} a_{\zeta} a_{\eta},
\]

where the vector of the model parameters \( w \) is

\[
w = [w_0, w_1, w_1, \ldots, w_{uv}, w_{uv}, \ldots] \mapsto [x_1, x_2, \ldots, x_m] \mapsto \sum_{j \in \mathcal{A}} w_j x_j,
\]

where the active set \( \mathcal{A} \) is a subset, which defines a regression model \( f_\mathcal{A} \) and \( w = [w_0, w_1, \ldots, w_{uv}] \) is a vector of regression parameters. Let the complexity \( \varepsilon f = |\mathcal{A}| \) of the model (1) be the number of its parameters.

Suppose the additive random variable \( \varepsilon \) of the regression model

\[
y = f(w, x) + \varepsilon
\]

to be normally distributed, \( \varepsilon \sim N(0, \sigma^2) \). If the random variable \( \varepsilon \) does not depend on \( x \), then the distribution of the target variable will be

\[
p(y|x, w, \beta, f) = \left( 2\pi \frac{1}{\beta} \right)^{-\frac{n}{2}} \exp \left( -\frac{1}{2} \sum_{i=1}^{n} \left( y_i - f(x_i) \right)^2 \right),
\]

where \( \beta^{-1} = \sigma^2 \) is the variance of \( \varepsilon \). According to the maximum likelihood estimation, the model quality criterion is the residual sum of squares

\[
S_X = \sum_{i \in X} \left( y_i - f(w, x_i) \right)^2,
\]

where \( X \subseteq I \) is some set of the object indices. This criterion is used in the model selection process as a part of the model evidence criterion (8), described below.

Also it is used as a part of the cross-validation technique in the comparative study: the data set is splitted into learning and test sample sets \( I = I_L \cup I_C \). Model selection algorithm defines the vector of parameters \( w \) of the model \( f \), which brings the optimal value to the quality criterion \( S_X \) using the leaning samples, \( \hat{\xi} \subseteq I_L \). The model of the optimal structure \( f \in \hat{\xi} \) minimizes the quality criterion \( S_C \) using \( X = I_C \).

3. Feature generation

Develop the following method to generate a set of features \( \{x_i\} \). Let \( \Xi = \{\xi_0\}, u = 1, \ldots, U \) be the set of non-generated variables. Let \( G = \{g_\varepsilon\}, v = 1, \ldots, V \), be the finite set of primitive functions given by experts (for example the set \( G = \{\sin, \cos, \exp, \sin \exp \} \)). Consider the Cartesian product \( G \times \Xi \). Each element \( (g_\varepsilon, \xi_0) \) of \( G \times \Xi \) is put in one-to-one correspondence with the superposition \( g_\varepsilon(\xi_0) \). Denote by \( a_\varepsilon \) the superposition \( g_\varepsilon(\xi_0) \), where \( \varepsilon = (v - 1)U + u \).

A polynomial in variables \( a_\varepsilon \) is the linear model (1) from the set \( \mathcal{F} \). Represent this polynomial in the form

\[
y = w_0 + \sum_{i=1}^{UV} w_i a_i + \sum_{i=1}^{UV} \sum_{\xi=1}^{UV} w_{ij} a_i a_{\xi} + \ldots + \sum_{i=1}^{UV} \sum_{\xi=1}^{UV} \sum_{\zeta=1}^{UV} \sum_{\eta=1}^{UV} \cdots w_{ij\xi\zeta\eta} a_i a_{\xi} a_{\zeta} a_{\eta},
\]

where the vector of the model parameters \( w \) is

\[
w = [w_0, w_1, w_1, \ldots, w_{uv}, w_{uv}, \ldots] \mapsto [x_1, x_2, \ldots, x_m] \mapsto \sum_{j \in \mathcal{A}} w_j x_j,
\]

where the active set \( \mathcal{A} \) is a subset, which defines a regression model \( f_\mathcal{A} \) and \( w = [w_0, w_1, \ldots, w_{uv}] \) is a vector of regression parameters. Let the complexity \( \varepsilon f = |\mathcal{A}| \) of the model (1) be the number of its parameters.
For further convenience normalize the generated features and the target vector:

\[
\sum_{i=1}^{n} x_i = 0, \quad \sum_{i=1}^{n} (x_i)^2 = 1, \quad \sum_{i=1}^{n} y^j = 0
\]

for any index \( j \in J \). Let all the generated vectors \( x_j = (x_j^1, \ldots, x_j^n) \) and \( x_k = (x_k^1, \ldots, x_k^n) \) are pair-vise linearly-independent for \( j, k \in J, j \neq k \). Exclude a linearly-dependent vectors from the generated set.

4. Model evidence

To estimate the model quality during the model selection process we use the coherent Bayesian inference and the notion of model evidence. Denote by \( p(f|D) \) the posterior probability of the model \( f \) given data set \( D \). According to the Bayes theorem it equals the product of the prior probability and evidence of the model over some normalization constant

\[
p(f|D) = \frac{p(D|f)p(f)}{p(D)}, \quad (6)
\]

where \( p(D|f) \) is the model evidence. Assume the prior probabilities of each model from the finite set model \( \mathcal{F} = \{f_1, \ldots, f_n\} \) to be equal. Under this assumption one can select the most probable model \( f \) using its evidence \( p(D|f) \).

Estimate the model evidence \( p(D|f) \) using the hypothesis on the distribution of the target vector (3). Since the regression problem is to find expectation of the target as a random variable, \( \mathbb{E}(y|x) = f(w, x) \), the hypothesis on the normal distribution of the target variable (3) implies the hypothesis on the normal distribution of the model parameters. If a parameter value \( w_j \) is equal to zero, it means the corresponding feature is not included in the model. Assume the expectation of the parameters equal to zero, \( \mathbb{E}(w) = 0 \). Then the distribution of the model parameters is

\[
p(w|\alpha, f) = \frac{1}{(2\pi\alpha^{-1})^{n/2}} \exp \left( -\frac{w^T w}{2\alpha} \right). \quad (7)
\]

Here, \( \alpha \) is the variance of the random variables \( w_j \) and the model parameters is a multivariate random variable \( w \) with independent items of the same variance.

According to the Bayes theorem the probability distribution of the model parameters depends on the data likelihood function (3) and the prior distribution of model parameters (7):

\[
p(w|D, \alpha, \beta, f) = \frac{p(D|w, \beta, f)p(w|\alpha, f)}{p(D|f)}.
\]

The denominator in the last expression is the model evidence in (6) and it is an integral

\[
\mathcal{E}(f) \overset{\text{def}}{=} p(D|f, \alpha, \beta) = \int p(D|w, \beta, f)p(w|\alpha, f)dw, \quad (8)
\]

Maximization of the integral could be replaced by the maximization of the integrand

\[
\mathcal{A} = \arg \max_{w \in \mathcal{F}} p(D|w, \beta, f) p(w|\alpha, f), \quad (9)
\]

The expressions (8) and (9) will be used in the model selection algorithm as the optimization criterion.

5. Multicollinearity detection

Since the described algorithm of feature generation may produce extremely large number of features, a multicollinearity problem arises during the model selection process. The proposed algorithm uses the Belsley method to detect multicollinear features. This method allows to point out the indices of highly correlated features.

To find these features compute the singular decomposition of the design matrix \( X \)

\[
X = U\Lambda V^T,
\]

where \( U, V \) are orthogonal matrices and \( \Lambda \) is a diagonal matrix with singular values \( \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \) on its diagonal. Since the singular values are non-negative and could be zero, remove the columns of the matrix \( X \), which corresponds to zero, so that \( \lambda_n > 0 \). Obtain the conditional indices of the matrix. These are defined as

\[
\eta_j = \frac{\lambda_1}{\lambda_j}.
\]

Note that each conditional index \( \eta_j \) corresponds to the feature \( x_j \).

Consider the variance inflation factor of \( j \)-th parameter of the model

\[
\text{VIF}_j = \frac{1}{1 - R^2_j}.
\]

The coefficient of determination \( R^2_j \) for \( j \)-th feature

\[
R^2_j = \frac{||x_j - X\alpha(\beta)w||^2}{||x_j - \bar{x}_j||^2},
\]

where \( || \cdot ||^2 \) is the Euclidian norm, and \( \bar{x}_j \) is the mean value for elements of the vector \( x_j \). The coefficient \( R^2_j \) refers to regression of the active features, but for \( j \)-th feature, to the \( j \)-th one. The variance of the parameter \( w_j \), an element of the parameter vector \( w \), is

\[
\sigma^2_{w_j} = \text{VIF}_j \frac{\sigma^2_{\alpha}}{||x_j - \bar{x}_j||^2}.
\]

Since the matrix \( X \) is invertible, then

\[
(X^T X)^{-1} = (V^T)^{-1} \Lambda^{-2} V^{-1} = V\Lambda^{-2} V^T
\]

and the variance of \( j \)-th parameter is

\[
\sigma^2_{w_j} = \sigma^2_{\alpha} (X^T X)^{-1} = \sigma^2_{\alpha} (V\Lambda^{-2} V^T).
\]

Denote by \( v_{ij} \) an element of the matrix \( V \) and represent the variance \( \sigma^2_{w_i} \) using the singular values of the matrix \( X \)

\[
\sigma^2_{w_i} = \sigma^2_{\alpha} \sum_{j=1}^{n} \frac{v_{ij}^2}{\sigma^2_{v_{ij}}} = \sigma^2_{\alpha} (q_1 + q_2 + \ldots + q_m). \quad (10)
\]

The large values of conditional indices \( \eta_j \) represent multicollinear dependencies in the features [9]. The large proportions of \( q_{ij} \) indicate the features to be removed.
6. Consequent model selection

Introduce two stages of the algorithm (see Figure 2). On the first stage add features one-by-one according to the model quality criterion (4). On the second stage delete the features one-by-one according to the multicollinearity criterion (10).

Let on the k-th step of the algorithm there is an active set $\mathcal{A}$, which defines the design matrix $X_\mathcal{A}$ so that $y = X_\mathcal{A}w$. At the zero step $\mathcal{A}_0 = \emptyset$. Describe the k-th step for “add” and “del” stages of the algorithm.

1. “Add”-stage. Append a feature number $j \in \mathcal{J} \setminus \mathcal{A}_{k-1}$ to the new active set

$$\mathcal{A}_k = \mathcal{A}_{k-1} \cup \{j\}$$

until the current evidence $E_k$ of the model $f_{\mathcal{A}_k}$ on the k-th step is not less than maximal value of the evidence during the add-stage minus some predefined constant $\Delta E$,

$$E_k \geq \max_{l \in [1, \ldots, L]} (E_{k-l}) - \Delta E. \quad (11)$$

Here $L$ is the number of steps passed during this stage. The index $j$ of the feature is defined by the append criterion. In the experiments described below the model quality (4) criterion was used:

$$j = \arg \min_{j \in \mathcal{J} \setminus \mathcal{A}_{k-1}} S_C(X_{\mathcal{A}_k}, y).$$

The criterion of LARS for this stage could be used as well.

2. “Del”-stage. Delete a feature number $j \in \mathcal{A}_{k-1}$ and form the new active set

$$\mathcal{A}_k = \mathcal{A}_{k-1} \setminus \{j\}$$

until the current evidence $E_k$ holds (11). The index $j$ of the feature is defined by the Belsley criterion (10),

$$j = \arg \max_{j \in \mathcal{A}} q_{i,j}, \quad \text{where } i = \arg \max_{j \in \mathcal{A}} \eta_i.$$

Repeat the stages until the evidence $E_k$ of the model is not empirically converged. The feature addition (deletion) stage holds on while the evidence is increasing. If the evidence of the current model starts to decrease it means that we obtain over trained (undertrained) model and it is time to change the stage.

![Evidence of the consequently generated models](image)

**Figure 2**: Evidence of the consequently generated models

Let on the $k$-th step the algorithm performs a random search minimizing the value of $S_C$, see [15]; GMDH, Group method for data handling [14]; Stepwise regression [7, 19, 23]; Ridge regression [7]; Tibshirani Lasso [16]; Stagewise regression [7]; FOS, Fast Orthogonal Search [17, 24]; LARS, Least angle regression [18]; Evidence

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>$S_L$</th>
<th>$S_C$</th>
<th>AIC</th>
<th>BIC</th>
<th>$C_p$</th>
<th>$\lg \kappa$</th>
<th>$k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genetic</td>
<td>0.073</td>
<td>0.107</td>
<td>-1152</td>
<td>-1072</td>
<td>3.37</td>
<td>13</td>
<td>25</td>
</tr>
<tr>
<td>GMDH</td>
<td>0.146</td>
<td>0.194</td>
<td>-1076</td>
<td>-1045</td>
<td>745</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>Stepwise</td>
<td>0.128</td>
<td>0.154</td>
<td>-1092</td>
<td>-1055</td>
<td>644</td>
<td>7</td>
<td>12</td>
</tr>
<tr>
<td>Ridge</td>
<td>0.111</td>
<td>0.146</td>
<td>-819</td>
<td>-330</td>
<td>832</td>
<td>33</td>
<td>160</td>
</tr>
<tr>
<td>Lasso</td>
<td>0.121</td>
<td>0.147</td>
<td>-1089</td>
<td>-1034</td>
<td>611</td>
<td>5</td>
<td>18</td>
</tr>
<tr>
<td>Stage</td>
<td>0.071</td>
<td>0.096</td>
<td>-1157</td>
<td>-1077</td>
<td>324</td>
<td>9</td>
<td>26</td>
</tr>
<tr>
<td>FOS</td>
<td>0.106</td>
<td>0.135</td>
<td>-1105</td>
<td>-1044</td>
<td>527</td>
<td>7</td>
<td>20</td>
</tr>
<tr>
<td>LARS</td>
<td>0.098</td>
<td>0.095</td>
<td>-1102</td>
<td>-1017</td>
<td>492</td>
<td>7</td>
<td>28</td>
</tr>
<tr>
<td>Evidence</td>
<td>0.097</td>
<td>0.123</td>
<td>-1118</td>
<td>-1054</td>
<td>469</td>
<td>5</td>
<td>21</td>
</tr>
</tbody>
</table>

7. Example and comparative study

Perform the comparative study for the proposed and well-known algorithms using historical data for the European stock option “Brent Crude Oil” (its ticker is CLG01, the underlying asset is NYM). Daily close prices were used for the numerical experiment. The history has 314 time ticks.

The data for this problem

$$\{(x^t, y^t) = \{(K_{i,t}, \sigma_i^t), i = 1, \ldots, m, \}$$

were obtained using historical data: option prices $C_{K,t}$ and underlying asset prices $P_t$, where $K \in \mathcal{K}$, $t \in \mathcal{T}$. For each sample $K_t$ and $t_i$ the value of implied volatility

$$\sigma_i^t = \arg \min_{C_{\text{hist}}(K_t, t_i)} (C_{\text{hist}}(K_t, t_i) - C_i(t_i)), \quad C_i(t_i) = \arg \min_{C_{\text{hist}}(K_t, t_i)} \left| \frac{\text{hist}(K_t, t_i) - C_i(t_i)}{C_i(t_i)} \right|$$

where the fair price of the option $C$ was calculated using the Black-Scholes formula [20].

The following primitive functions were used: $G = [x^{-1}, x^0, \sqrt{x}, x, \ln(x), \tan(x)]$. The degree of the polynomial (5) equals three. The data set was split to the training $I_L$ and test $I_G$ subsets of equal number of elements.

The results of the comparative study are shown on the table 1. The table has the following columns:

- $S_L$ and $S_C$, see (4), were averaged on 10 different splits of the data set;
- AIC — Akaike Information Criterion, $AIC = m \ln \frac{\sum_i y_i}{m} + 2|\mathcal{A}|$, see [3];
- BIC — Bayesian Information Criterion, $BIC = m \ln \frac{\sum_i y_i}{m} + |\mathcal{A}| \ln m$, see [21];
- $C_p$ — Mallows criterion, $C_p = \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} - m + 2|\mathcal{A}|$, see [22];
- $\lg \kappa$ — logarithm to base 10 of the design matrix $X_{\mathcal{A}}$ singular values’ ratio, $\kappa = \frac{\lambda_{\max}}{\lambda_{\min}}$;
- $c_f = |\mathcal{A}|$ — the number of features in the design matrix $X_{\mathcal{A}}$.

The following algorithms were used in the study: Genetic algorithm performs a random search minimizing the value of $S_C$, see [15]; GMDH, Group method for data handling [14]; Stepwise regression [7, 19, 23]; Ridge regression [7]; Tibshirani Lasso [16]; Stagewise regression [7]; FOS, Fast Orthogonal Search [17, 24]; LARS, Least angle regression [18]; Evidence
maximization, the proposed algorithm. The results of the comparative study are shown in the Table 1.

Figure 3 shows how the evidence of a model typically depends on the model complexity. The x-axis shows the model complexity \( c_f \). The y-axis shows the evidence of a model. Each blue dot represents a generated model. The maximum evidence corresponds to the design matrix of 6 features.

8. Summary

The combination of two strategies: feature generation and model selection could bring good results, especially if the linear combination of original features can not be used as an adequate model. However in case of massive feature generation, many highly correlated features appear as the result of generation. To reduce the possibility of this phenomenon we suggested to use multicollinearity detection and evidence maximization methods in the stepwise algorithm. The comparative study shows that in this case we could obtain well-conditioned design matrix for the linear model and fair square sum of errors at the same time.

References