Regular Paper

An Incremental Approach for Optimal Feature Selection in Regression: A Case Study of Wagyu Proteome Analysis

Nanami Higashiguchi†, Masatsugu Motohiro†, Haruka Ikegami*,
Tamako Matsuhashi*, Kazuya Matsumoto*, and Takuya Yoshihiro‡

†Graduate School of Systems Engineering, Wakayama University, Japan
*Graduate School of Biology-Oriented Science and Technology, Kindai University, Japan
‡Faculty of Systems Engineering, Wakayama University, Japan
‡tac@wakayama-u.ac.jp

Abstract - Sparse modeling has attracted significant attention as big data analysis goes popular. LASSO is one of the sparse modeling techniques to retrieve a set of features correlated to a target function. LASSO runs in very low computational time to obtain near-optimal set of features even if the data set is very large. However, due to its own regularization term, optimization errors are not negligible. In several practical scenes, it is required to obtain more optimal solutions within feasible computational time. However, the solution has not been presented clearly. In this paper, we present a new heuristic approach called IFS (Incremental Feature Selection) that starts from a collection of singleton feature sets, and increases features in a set one by one to finally obtain a quasi-optimal M-element feature set. The proposed technique IFS is applied to the analysis of Wagyu proteome expression data set, and we proved that IFS performs better than LASSO. Simultaneously we introduce a technique to find commonly-correlated features for the same objective function among multiple groups of data sets. We can use Multi-task LASSO for this purpose, but since it does not aware of the uniformity of the effects on each group, it is not enough to identify commonly correlated feature sets among all the groups. Our technique in IFS uses a fairness index to tackle the problem. We applied IFS to the Wagyu data set and showed that IFS ensures to retrieve a quasi-optimal feature set whose fairness index among correlation of those groups is larger than the given threshold.

Keywords: Sparse Analysis, LASSO, Feature Selection, Wagyu, Proteomics

1 INTRODUCTION

Sparse modeling has attracted significant attention in face of big data analysis. Recently, large dimensional data sets are easily obtained such as biological data represented by gene or protein expression profiles, and are recognized as very useful sources to analyze valuable properties of creatures. However, in computational analysis with those data, variable selection that retrieves a variable set that highly correlates the target trait from a vast amount of features is an essential task to pursuit. With the naive execution, the computational time of this task explodes to exponential and usually not feasibly solved with the current computers. To solve the task within feasible time, LASSO (Least Absolute Shrinkage and Selection Operator) [1], which retrieves a quasi-optimal variable set that minimizes the square errors in multiple regression analysis, is one of the well-recognized feature selection methods from vast amount of features included in the original data set. LASSO applies L1-norm regularization term in optimization formula to obtain a quasi-optimal feature set within feasible time. However, LASSO has a problem that the obtained feature set in many cases has a considerably large error and sometime far from optimal. To obtain a feature set closer to the optimal within feasible time is one of the solicited research tasks in this field.

In this study, we tackle this problem with a case study of Wagyu analysis, in which we try to find a small feature set from hundreds of proteins in a given protein profile that significantly correlates with Wagyu beef quality. Additionally, because the data set includes samples (i.e., beef cattle) from multiple Wagyu regions, we try to separate the common trends and each regional trend. We propose a new method for these tasks to treat a data set with hundreds of features.

Wagyu is known as a high-quality branded beef of Japan, with a feature of soft and tender meats due to fats mixed in the meat. There are several regions famous for Wagyu in Japan, and each region has different policy of breeding sires and beef cattle to produce larger amount of higher-quality meat. This contention among regions has improved the methodology of breeding beef cattle so far. However, since they mostly depend on traditional methods based on statistics on bloodlines or breeding experience, there is an apparent limitation in beef-quality improvement.

Recently, several comprehensive analyses in genomics or proteomics have been developed, for example, gene and protein expression profiles that include expression values of so many genes and proteins are available with smaller cost than ever. Specifically, we have a large number of explanation variables retrieved from each sample, which potentially makes us predict beef quality of each beef cattle in the early stage of beef-cattle breeding. This also could lead to the innovative methodology of breeding beef cattle to improve its beef quality.

Here, the first problem is that the variables in genes or protein profiles are so many that we can hardly select the optimal variable set to predict beef quality. The second prob-

lem is that beef cattle of distinct regions has different trend on its data so some analytic methods to treat this problem is required. As for the first problem, recently sparse analyses have been developed in which near optimal feature selection is possible with small computational cost. Especially, if we intend to perform multiple regression, LASSO is often used. LASSO minimizes MSEs (Mean Square Errors) in the form of multiple linear regression, in which by using L1 regularizer most of the coefficients are to shrink to zero. LASSO actually selects a near-optimal variable set within feasible time even if the number of available variables is very large. However, LASSO has a problem in Wagyu analysis that it cannot catch up with the trend of each branded Wagyu regions.

Multi-task LASSO(MT-LASSO) [2], which considers multiple objective functions in selecting a variable set has been proposed. MT-LASSO applies L1/L2 penalty to retrieve a variable set that commonly explains the multiple objective functions. This by definition can be used to explain the trend of each region of Wagyu brand by retrieving the common variables that explain trends of the all target regions. However, MT-LASSO retrieves a variable set without considering the balance of effects among multiple regions so that it may select a variable set that strongly effects on a region while weakly effects on other regions. Furthermore, it is known in both LASSO and MT-LASSO that the selected variables are not always optimal in terms of multiple regression so that we can hardly retrieve the optimal set of variables that explains the target traits of Wagyu beef [3]. Methods to retrieve the optimal variable set while considering multiple Wagyu brand regions are required.

In this paper, we present a solution for this problem, i.e., we propose a variable selection method IFS (Incremental Feature Selection) that retrieves an optimal commonly effecting variables among multiple Wagyu regions within a feasible computational time. We first exploit a single regression results, i.e., correlation coefficients, and fairness indices among them to retrieve a small number of variables as a candidate of selected variables. Second, we make a pair of those selected variables, and retrieve a certain number of pairs from them using the multiple correlation coefficients and the fairness indices among regions. We repeat this process to make candidate combinations including a larger number of variables. By testing all combinations of the candidate variables with multiple regression and fairness indices, we finally retrieve the best variable set within feasible time.

This paper is organized as follows. In Section 2, we describe the trend of Branded Wagyu beef. In Section 3, we introduce LASSO and MT-LASSO. In Section 4, we present a proposed method and its algorithms, and its computational complexity is analyzed in Section 5. After the evaluation results shown in Section 6, finally we conclude the work in Section 7.

2 BRAND WAGYU BEEF

Japanese Black Cattle is a beef cattle peculiar to Japan, which produces various brand beef called Wagyu such as Kobe beef, etc. There are many regional brand beefs in Japan, each of which has its own way to breed cattle, and apply its own criterion to authorize whether each head of cattle is sold under the name of the brand beef. As the authorization criteria, there are several items, e.g., the birth of cattle, the way to raise cattle, the rating of beef, etc. Among them, the rating of beef is the most important. The rating criteria include various values, and especially 6 items among them are regarded as the most important ones to judge whether a head of cattle is authorized as brand beef [5]. The 6 items, which we call *economical traits*, are CW (Carcass Weight), BMS (Beef Marbling Standard), YE (Yield Enhancement), RT (Rib Thickness), SFT (Subcutaneous Fat Thickness), and REA (Rib-Eye Area). Basically from these criteria, the price of beef in the market is determined. Therefore, the farmers of brand beef have been made a great endeavor to produce quality beef.

Wagyu farmers take various methodologies to produce quality beef stably. One of the most important methods is to control bloodline so as to have better values of the economical traits. Since the bloodline is known to have close relationship with economical traits, efficient inbreeding by producing and identifying genetically excellent individual cattle has a significant importance to improve the value of brand beef site. Each brand-beef site usually breeds several head of cattle called sires that have excellent genetic ability [6], [7]. From sires, we take sperms and freeze them, and sell them to farmers. With this system, excellent bloodline of sires is distributed to farmers and generate thousands of children cattle from an excellent sire. Note that, in brand-beef sites, father of each beef cattle is called '1-generation ancestor' and the father of beef cattle is one of the most important criteria to predict economical traits of beef cattle.

As a statistical methodology to predict economical traits of beef cattle from past records, the breeding values are usually used in brand beef sites. The breeding values are calculated for each economical trait, which represent the ability to improve the trait values compared to the average ability in the group. There are two kinds of breeding values, i.e., estimated breeding values and expected breeding values. The former is calculated for sires who have descendants with carcass characteristic scores and represents the ability to improve 6 economical traits. In contrast, the latter is calculated for each beef cattle that does not have enough number of descendants to estimate breeding values.

We have several variations of bloodline models used to compute breeding values. Currently, the most frequently used model is so called 'animal model,' which considers all the relative relationship including brothers of beef cattle that have the same mother. With a bloodline model and the data set, BLUP method calculates the breeding values in a statistical manner as the genetic ability inherited through bloodlines [8]. The expected breeding value for each beef cattle is calculated as the average of its two parents.

On the other side, raising method to produce high-value beef cattle stably also has been studied so far. However, methods in this area are mostly depends on experiences of farmers, and are not based on any scientific results or real data. For example, livestock associations or stock farmers have accumulated their experience to raise high-value beef cattle as know-how or some kind of manuals. This kind of information

has wide variations from direct methods such as how to feed cattle to indirect methods such as the structure of cowsheds. As for the academic results, a few studies have been published on the relationship between raising methodology and economical traits. For example, there is a study on improving BMS values by controlling the concentration of vitamin A [9]. However, in the current state, we have still too little knowledge to actually control economical traits in raising in livestock farms.

3 LASSO AND MULTI-TASK LASSO

LASSO [1] is a well-known technique for feature selection from the large number of features based on linier regression models. Let S be the given set of samples, and F be that of features. Let $x_{sf}(s = 1, 2, \ldots, |S|, f = 1, 2, \ldots, |F|)$ be the measured feature value of sample s on feature f, where $|S|$ and $|F|$ are the number of elements of S and F, respectively. Hereafter we may write just S and F in place of $|S|$ and $|F|$ for conciseness. Let $\mathbf{x}_s = (x_{s1}, x_{s2}, \ldots, x_{sF})^\mathsf{T}$, be the measured vector for each sample $s \in S$, where A^T denotes a transposed matrix of A. Let $X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_S]$ be the matrix of the feature data. Let y_s be the measured trait values for each sample s, and $\mathbf{y} = (y_1, y_2, \ldots, y_S)$ be the trait vector. Then, LASSO is formulated as follows:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}}(\|\mathbf{y} - \boldsymbol{\beta}\mathbf{X}\| + \lambda|\boldsymbol{\beta}|) \tag{1}$$

where λ is a non-negative regularization parameter, and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_F)$ is a coefficient vector for X. Additionally, $\|\boldsymbol{\beta}\|$ represents the L2 norm of a vector $\boldsymbol{\beta}$ defined as $\|\boldsymbol{\beta}\| = \sqrt{\sum_{f\in F}\beta_f^2}$, and $|\boldsymbol{\beta}|$ represents the L1 norm defined as $|\boldsymbol{\beta}| = \sum_{f\in F}|\beta_f|$ Due to the effect of L1-norm penalty with λ, most of β_f converges to zero during the computation of the optimal solution. As a result, we have a small number of non-zero coefficients, and this process works as a feature selection from a large number of feature variables.

Multi-task LASSO [2] is an extension of LASSO, which treats multiple objective functions. Let us denote T as the set of tasks (i.e., set of objective functions), and also let us define a feature matrix and a trait vector for each task. Namely, we let $x_{sf}^{(t)}$ be the measured feature values for task $t \in T$. Also, let $y_s^{(t)}$ be the measured trait values for task $t \in T$. Similarly, we also write $\mathbf{x}_f^{(t)}$, $\mathbf{X}^{(t)}$, $\mathbf{y}^{(t)}$, $\boldsymbol{\beta}^{(t)}$ etc. Note that the number of samples for each task $S^{(t)}$ could be different. Then, the MT-LASSO is expressed as follows:

$$\hat{\boldsymbol{W}} = \arg\min_{\boldsymbol{W}}(\sum_{t\in T}\|\mathbf{y}^{(t)} - \boldsymbol{\beta}^{(t)}\mathbf{X}^{(t)}\| + \lambda|\mathbf{w}|), \tag{2}$$

where W is the coefficient matrix that combines all coefficient vectors, defined as $W = [\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, \ldots, \boldsymbol{\beta}^{(T)}]$, and $\mathbf{w}$ is the vector of L2 norms of the coefficients where $\mathbf{w} = (\|\mathbf{w}_1\|, \|\mathbf{w}_2\|, \ldots, \|\mathbf{w}_F\|)$, and $\mathbf{w}_f$ for $f \in F$ is defined as $\mathbf{w}_f = (\beta_f^{(1)}, \beta_f^{(2)}, \ldots, \beta_f^{(T)})$. Note that the regularization term is the combination of L1 and L2 norms, as shown in Fig. 1. In MT-LASSO, the coefficients $\beta_f^{(t)}$ are defined for each $f \in F$ and $t \in T$. To proceed feature selection and coefficients optimization altogether, MT-LASSO uses the combination of L1/L2 regularization. First, L2-norm of w_f, the coefficients vector of the same feature is computed, and second, L1-norm of those are used in the regilarization term. This enables us to select the commonly effective features for all tasks first, and then to optimize coefficients of the selected features within each task.

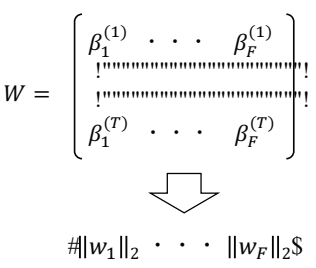


Figure 1: L1/L2 Regularization in Multi-task LASSO

MT-LASSO can be applied to our problem. Note that, in each region t of brand Wagyu, distinct samples, i.e., heads of beef cattle, are grown up so that we have measured feature sets $\mathbf{X}^{(t)}$ for each region t. As for the trait, we apply the same trait such as BMS, but each region has their own beef cattle, so that the data is expressed as $\mathbf{y}^{(t)}$. By solving MT-LASSO with the above $\mathbf{X}^{(t)}$ and $\mathbf{y}^{(t)}$, we can obtain the commonly effective feature set among multiple regions within the framework of MT-LASSO. However, the problem is that MT-LASSO does not consider the balance of effects among multiple regions. Additionally, MT-LASSO lacks optimality so that a non-optimal set of features would be selected frequently. In this paper, we try to improve the optimality utilizing correlation coefficients between $\mathbf{y}^{(t)}$ and $\mathbf{x}_f^{(t)}$ while balancing the effects on multiple regions using a fairness index.

4 THE PROPOSED METHOD

4.1 Problem Formulation

In this study, with a given set of regions T, we retrieve a set of features that has comparably high correlation for all region $t \in T$. For each region $t \in T$, for a given measured trait set $\mathbf{y}^{(t)}$, and a measured feature set $\mathbf{X}^{(t)}$ for features F, we try to find a feature set $F' \subseteq F$ that minimizes total MSE under the constraint that the fairness index of MSEs among all regions is larger than a given threshold. We use Jain's fairness index [4] to measure the 'fairness,' namely, to measure the uniformity in the effect of those feature set. This index takes one when all the values are the same, and takes n^{-1} in the worst case, where n is the number of input values (i.e., regions in this study). In the proposed method, the fairness in MSE among regions are defined as follows.

$$FI_{F'} = J(E_{F'}^{(1)}, \ldots, E_{F'}^{(T)}) = \frac{(\sum_{t\in T} E_{F'}^{(t)})^2}{n\sum_{t\in T}(E_{F'}^{(t)})^2} \tag{3}$$

Here, for the retrieved feature set F', we let $E_{F'}^{(t)}$ be the

MSE in region $t \in T$, and let $FI_{F'}$ be the fairness index of MSEs among all regions. Also, we let $E^{(all)}_{F'}$ be the MSE computed from all samples s in all regions T. Then, the problem formulation to solve in this paper is shown as follows.

Problem Formulation

Given M, the number of features to retrieve, and J, the least required value of $FI_{F'}$, find the feature set F' that minimizes $E^{(all)}_{F'}$ under constraints $|F'| = M$ and $FI_{F'} \geq l$.

4.2 Proposed Algorithm

As shown above, we propose an algorithm to compute a feature set that marks equally high correlation in every region $t \in T$. Specifically, since LASSO minimizes MSE, we also use MSE as the performance index, and compute a feature set that takes equally small MSEs for the given feature values $\mathbf{x}^{(t)}$ for each region $(t \in T)$. Generally, the computational complexity explodes when we select a set of M features that leads minimum MSE from a large number of features because we must compute MSEs through multiple regression for every combination of M features in the data set. In our algorithm, we solve this problem by increasing the number of features step by step. Namely, we start from a set of single features, next we make pairs of features, and then triads of features, and so on. Every time we increase the member of the sets, we filter the sets to limit the number of sets in order to limit the computational time. Generally, in multiple regression analysis, MSE values for a set of features tends to be smaller when a part of them leads to low MSE values. By making use of this property, we repeat increasing the member of features in the combinations and filter them, and finally obtain the best set of M features within feasible computational time.

Specifically, to retrieve several features out of hundreds or thousands of features, we first make a single regression analysis, compute MSEs, and filter them with those MSEs. With the reduced number of features, we make all possible pairs of the features, and filter them to retrieve pairs that have uniformly high MSEs in multiple regression analysis. Next, we make all possible triads of features by combinatorially adding one feature to the pairs, and filter them in the same way to retrieve a feasible number of triads. By repeating those, we finally obtain a set of M features that has good MSEs as well as good uniformity in MSEs among regions. By limiting the number of feature sets in each stage, we provide a guarantee on computational time to be feasible.

As aforementioned, we denote the given regions of Wagyu brand by $t \in T$, feature set by F, measured value vector for feature $f \in F$ by $\mathbf{x}^{(t)}_f$, measured value matrix for all features by $\mathbf{X}^{(t)}$, and measured trait vector by $\mathbf{y}^{(t)}$. For each stage i of our algorithm, G_i is input and G_{i+1} is output, where G_i is a family of feature sets represented by $G_i = \{F_{1,i}, F_{2,i}, \ldots, F_{k,i}\}, (1 \leq k \leq N)$, in which $F_{k,i} \subseteq F$, $|F_{k,i}| = i$, and N is a predefined natural number.

We present the algorithm to obtain the solution F' in the following.

1. Initialize with $i = 1$ and $G_i = F$.
2. Compute MSEs by applying multiple regression analysis between $\mathbf{y}^{(t)}$ and $F_{k,i} \in G_i$, and get $E^{(t)}_{F_{k,i}}$ for each $t \in T$ and $F_{k,i} \in G_i$. If $i = 1$, since the number of features in $F_{k,i}$ is one, we apply single regression analysis between each feature $f \in F_{k,i}$ and $\mathbf{y}^{(t)}$ instead.
3. Compute the fairness index $FI_{F_{k,i}}$ from $E^{(t)}_{F_{k,i}}$ $(^\forall t \in T)$ for each $F_{k,i} \in G_i$.
4. Compute MSEs with all samples of all regions, i.e., compute $E^{(all)}_{F_{k,i}}$ for each feature set in $F_{k,i}$.
5. Obtain a family of feature sets $G'_i \subseteq G_i$ by retrieving feature sets $F_{k,i}$ with $FI_{F_{k,i}} \geq Jl$.
6. If $|G'_i| > N$, limit the number of elements in G_i: Obtain $G''_i \subseteq G'_i$ by retrieving N-smallest feature sets in G'_i with respect to $E^{(all)}_{F_{k,i}}$. Otherwise, $G''_i = G'_i$.
7. Obtain the family of feature sets G_{i+1} as follows.
   (a) Create a set of features included in G''_i. Specifically, retrieve all features included in some feature sets in G''_i, and make a feature set $H_i = \{f_{1,i}, f_{2,i}, \ldots, f_{n,i}\}$.
   (b) Create a family of feature sets G_{i+1} by making all combinations between $G''_i = \{F_{1,i}, F_{2,i}, \ldots, F_{k,i}\}$ and $H_i = \{f_{1,i}, f_{2,i}, \ldots, f_{n,i}\}$. Namely, $G_{i+1} = \{F_{1,i} \cup \{f_{1,i}\}, F_{1,i} \cup \{f_{2,i}\}, \ldots, F_{1,i} \cup \{f_{n,i}\}, F_{2,i} \cup \{f_{1,i}\}, \ldots, F_{2,i} \cup \{f_{n,i}\}, \ldots, F_{N,i} \cup \{f_{1,i}\}, \ldots, F_{N,i} \cup \{f_{n,i}\}\}$.
   (c) Remove duplicated elements in G_{i+1}.
8. If $i < M$, do $i = i + 1$ and return to step 2.
9. Select the least MSE feature set F' that satisfies $FI_{F'} \geq J$ from G''_M, and output it.

We explain each steps of the algorithm. See Table.1 for definitions of variables.

First of all, in Step 1, we initialize variables i and G_i.

In Step 2, we compute MSEs by applying multiple regression for all feature sets in G_i and the target trait. Specifically, for each region $t \in T$ and feature set $F_{k,i} \in G_i$, we perform multiple regression analysis between $x^{(t)}_{F_{k,i}}$ and $\mathbf{y}^{(t)}$, and obtain the MSE value $E^{(t)}_{F_{k,i}}$ as a result.

In Step 3, we compute the fairness index for each feature set in G_i based on the formula (3), which represents the uniformity of the effects of $F_{k,i}$ on each region.

In Step 4, we apply multiple regression and compute MSE for each $F_{k,i}$ again, but using all samples $S^{(t)}$ in all regions in T altogether. We let $\mathbf{x}^{(all)}_F = (\mathbf{x}^{(1)}_F, \mathbf{x}^{(2)}_F, \ldots, \mathbf{x}^{(s)}_F)$ be the feature matrix and let $\mathbf{y}^{(all)} = (\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \ldots, \mathbf{y}^{(s)})$ be the trait vector. Then we can compute MSE denoted by $E^{(all)}_F$ as a result of multiple regression of $\mathbf{x}^{(all)}_F$ and $\mathbf{y}^{(all)}$. Note that $E^{(all)}_{F_{k,i}}$ is computed for each $F_{k,i} \in G_i$.

Table 1: Notations

Symbol	Description
i	Current processing stage.
F	Set of Features.
T	Set of Wagyu region t
$\mathbf{x}_f^{(t)}$	Measurement vector of feature f in region t.
$\mathbf{y}^{(t)}$	Measurement vector of a trait with region t.
G_i	Family of feature sets in i-th stage.
H_i	Set of features included in G_i.
$E_{F_{k,i}}^{(t)}$	MSE value in i-th stage computed from regression with $x_F^{(t)}$ and $\mathbf{y}^{(t)}$ in region t.
$E_{F_{k,i}}^{(all)}$	The MSE in i-th stage computed with all samples in all regions.
$FI_{F_{k,i}}$	Jain's fairness index computed for $F_{k,i}$.
J	Threshold on fairness index given as constraint for output.
l	Margin for J to allow possible candidates with smaller value than J. In algorithm, we apply Jl as the threshold.
N	The number of feature sets to be selected on each stage.
M	The number of features to be selected finally.

In Step 5, we filter the feature set in G_i using Jain's fairness index to exclude the feature set unlikely to be a candidate for final output. With Jain's fairness index, we examine the uniformity of MSEs for each region $t \in T$. If the uniformity is high, the feature set is said to have correlation equally to all regions, meaning that the feature set includes general effect for Wagyu, not depending on regions. Through preliminary test, we found that the MSEs and fairness indices computed with a feature set F has small difference from those computed from the feature set $F' = F - \{f\}$ for $f \in F$. Thus, our basic strategy is to keep feature sets whose fairness indices are high enough in each stage of the algorithm. We apply threshold $J \times l$ to $FI_{F_{k,i}}$ and obtain the feature set $G_i' \subset G_i$. The threshold J and l are predefined constants, where J represents the requirement on fairness index for the final output of the algorithm. l provides a mergin for threshold J to allow feature sets with a little smaller value than J included in a candidate feature set G_i. Note that a feature set in stage i that has a little smaller fairness index than J can increase its fairness index in stage $i+1$ by adding one feature. The margin l works to keep the potential candidates for the next stage.

In Step 6, we limit the number of feature set in G_i' up to N elements. This step aims at ensuring the computational time to be feasible. As we mention later, the computational time highly depends on N. This is done by using MSE values $E_{F_{k,i}}^{(all)}$ computed in Step 4, i.e., if $|G_i'| > N$, we select top-N feature sets in terms of MSE, and create G_i''. Note that MSE is the most important selection criteria in the objective of this paper, and in this paper, we intend to obtain the minimum MSE feature set under the constraint that the fairness index is larger than J.

In Step 7, we create a family of feature sets for the next stage, i.e., G_{i+1}. When we increment the processing stage from i to $i+1$, the number of features in the feature set also incremented by one. Our basic strategy is to make all combinations of adding one feature to each feature set $F \in G_i''$. As candidate features to add, we make a feature set H_i that consists of all features included in the feature set family G_i'', and make all combinations of H_i and G_i''. By removing duplicated feature sets, we regard the set as G_{i+1}.

In Step 8, we repeat the above process until the number of elements in the feature set reaches M.

Finally in Step 9, we select the best feature set from G_M'' and output it. The best feature set is the one that have minimum MSE value among the sets whose fairness indices are larger than or equal to J.

5 COMPUTATIONAL COMPLEXITY

In this section, we analyze the computational complexity of the algorithm. The computational complexity depends on S, the number of all samples in all regions, N, the maximum number of feature sets in each stage, M, the number of stages, and T, the number of regions. The initialization process in Step 1 is clearly $O(1)$. In Step 2, we compute MSE for each feature set and for each region. Since the number of feature sets is less than N^2, and the complexity of multiple regression is $O(S)$, the complexity of this part is $O(SN^2)$. In Step 3, we compute fairness index for each feature set. The complexity to compute a fairness index is $O(T)$ when MSE values for each region is given. Thus, the complexity of this part if $O(TN^2)$. Since $T \ll S$, this can be regarded as $O(N^2)$. In Step 4, we do multiple regression with all samples in all regions for each feature sets, taking $O(SN^2)$ time. In Step 5, we remove feature sets whose fairness indices are less than Jl from G_i, which takes $O(N^2)$ time. In Step 6, we sort the feature sets by MSE, and select top-N sets. Since we sort at most N^2 elements, the complexity is $O(N^2 logN^2) = O(N^2 logN)$. In Step 7, we create the family of feature sets G_{i+1}. Since the number of elements is at most N^2, the complexity is $O(N^2)$. Step 8 and 9 apparently takes $O(1)$ time. The above is the complexity for one stage. Since we have M stages, the total complexity is $O(SMN^2 logN)$.

In Fig. 2, we show the real computational time in our evaluation described later. The parameters are $S = 96, M = 6$, and $N = (50, 100, 150, 200, 250, 300, 350, 400, 450, 500, 550)$, and the proposed method is executed on a general personal computer equipped with Intel i5 2.8GHz CPU and 8GB Memory. The results show that the computational time increases as N increases, but the slope is not steep. Although the number of samples in this data set is not large, the computational time of the proposed algorithm is feasible for a certain magnitude of data size.

6 EVALUATION

6.1 Data Description

We evaluate the proposed method compared with the result of Multi-task LASSO (MT-LASSO). The data consists of 3

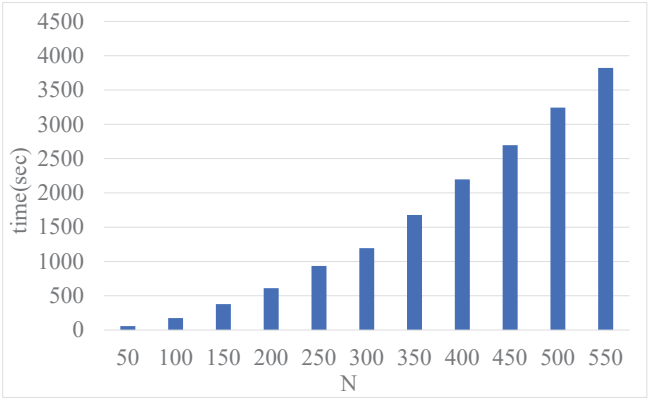


Figure 2: Execution Time of Proposed Algorithm

regions of branded Wagyu, which we refer region A, B, and C, and each region has 51, 10, and 35 beef cattle (i.e., samples) in the data. Each beef cattle has been grown up in one of the regions, and the 6 economical traits were measured before slaughtered and sold as meat. As aforementioned, the 6 economical traits are CW (Carcass Weight), REA (Rib-Eye Area), RT (Rib Thickness), SFT (Subcutaneous Fat Thickness), YE (Yield Enhancement), and BMS (Beef Merbling Standard). As a result, our data has 6 trait values for each sample from 3 regions.

The feature data set is a proteome expression profile of serum; for each beef cattle, serum is taken with the interval of 3-4 months, which are analyzed by SWATH-MS [10] (Sequential Window Acquisition of all Theoretical fragment ion spectra Mass Spectrometry) method with our own preprocessing treatment. In this method, we got the expression levels of 135 proteins for each sample. As a result, we got 135 protein expression values for 6 periods of time, so we have $135 \times 6 = 810$ features for each sample from 3 regions. After removing the features that contain null values, we have 580 features to apply the proposed method.

6.2 Evaluation Methods

We applied the proposed method and MT-LASSO to the data set described above. MT-LASSO originally is a feature selection method based on multiple objective functions, but it can be applied to our problem, i.e., it treats the same objective function for different data groups. To the best of our knowledge, MT-LASSO is the only method that treats multiple regions under a single objective function.

As evaluation criteria, we use MSE and the fairness index to compare those two methods. We select MSE rather than correlation coefficients to compare performance of the two because LASSO (as well as MT-LASSO) is an optimization scheme based on MSE. We also use Jain's fairness index [4] to measure the uniformity of the effects (i.e., correlation measured by MSE) among multiple Wagyu regions. When the fairness index takes high value, i.e., close to 1, we can regard that the selected feature set has uniformly the same level of correlation in all regions, meaning that the effect is not specific in a particular region, but expresses a general property in Wagyu. To separate the general effect from regional ones is the objective of this study.

As parameters, we set $M = 6$, i.e., we retrieve a set of 6 features to explain target traits. We choose this value considering the number of samples in each region. We also set $J = 0.8$, $l = 0.8$, and $N = (50, 100, 150, 200, 250, 300, 350, 400, 450, 500, 550)$, which are determined through preliminary tests. In MT-LASSO, to compare the performance with the proposed method, we adjust the parameter value λ to select exactly 6 features, and used the results in our comparison. We tried to use MT-LASSO implementation included in scikit-learn [12], but unfortunately, this cannot treat our dataset; it does not support the case with the same objective function applied to multiple groups. However, when we focus on the feature selection function, the mechanism of MT-LASSO is exactly the same as LASSO. (Notice that MT-LASSO first makes a feature selection based on L1 norm penalty, and then determine coefficients for each objective function based on L2 norm.) Thus, we apply LASSO implementation included in Python scikit-learn library for the results of MT-LASSO. Additionally, to expect the fairness in comparison, we do not use the MSE value obtained directly from MT-LASSO. To get rid of the effect of penalty terms, we made a multiple regression analysis with the features retrieved by MT-LASSO, and used the MSE values in our comparison.

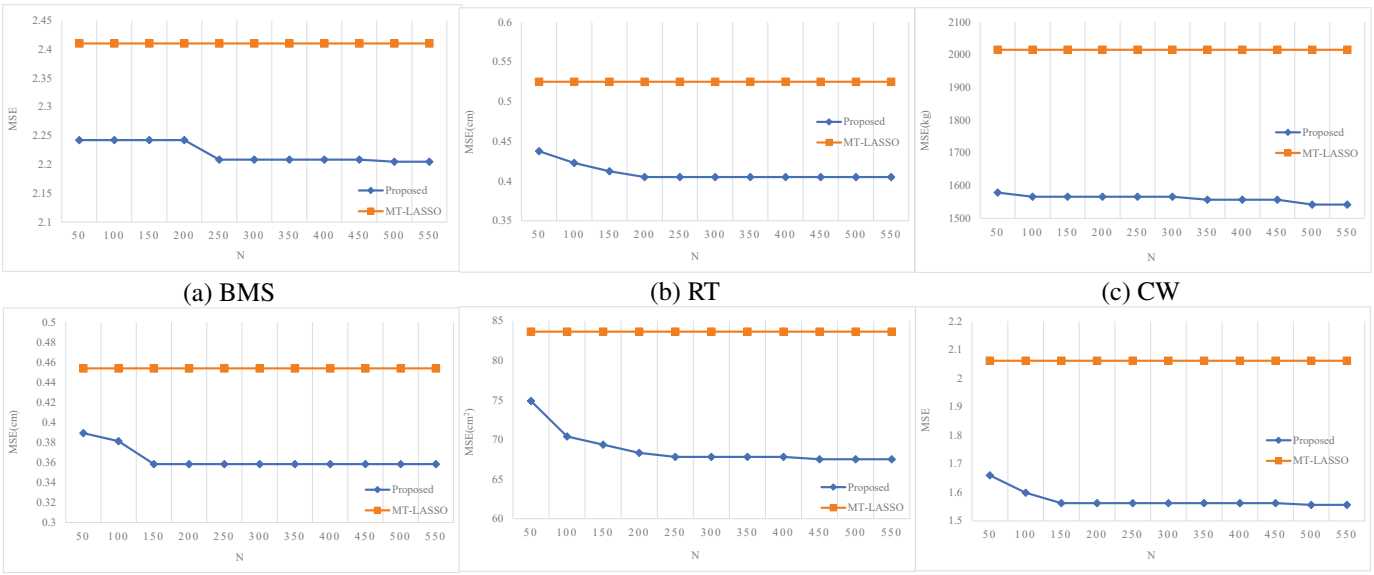
6.3 Evaluation Results

In Fig. 3, we show the comparison results on MSE values for each target trait. We see that the proposed method outperforms MT-LASSO in every trait. The cause of errors in MT-LASSO is the L1-norm regularization term. In contrast, the proposed method selects feature sets in the stepwise manner during which we keep good combinations of features as candidates of the solution. These results show that our strategy clearly works, and the performance is better than MT-LASSO (as well as LASSO) even if we keep only 50 ($N = 50$) candidates in each stage. Figure3 also shows that the performance goes better when N is increased, although the performance improvement is not significant. Note that, in this figure, MT-LASSO is shown to always take the same value since it does not have parameter N.

In Fig. 4, we show the results on fairness index. We see that, in most cases, the proposed method presents better performance than MT-LASSO, meaning that the proposed method surely retrieves feature sets that uniformly effects on all regions. We see that the performance of the proposed method involves some fluctuation, which is not seen in MSE performance. This is because the primal objective function is not fairness index, but MSE. It is natural that the best-MSE feature sets do not always have the best fairness index value.

More importantly, we see that the proposed method always keeps the constraint of fairness index, i.e., the value must be larger than $J = 0.8$. Recall that the proposed method intends to minimize MSE under the constraint on fairness index. We succeeded to keep the constraint and to ensure that the fairness index is above the preconfigured threshold $J = 0.8$.

In summary, we showed that the proposed method not only ensures the minimum bound on fairness index, but also mark better optimality on MSE compared with MT-LASSO as well as LASSO.



(a) BMS (b) RT (c) CW

(d) SFT (e) REA (f) YE

Figure 3: Results on MSE

7 DISCUSSION ON APPLICATION

In this section, we discuss how to apply the proposed methods in practice. We proposed two different techniques in this paper. One is IFS, which takes the incremental approach to select optimal set of features, and the other is an extension of IFS, which enable us to retrieve a commonly effective feature set among multiple groups. The first method IFS solves the same problem as well-known LASSO, which is used widely in variation of feature selection cases. Thus, IFS is also applicable a wide variety of practical cases to obtain the optimal set.

The second method, the extension of IFS, is compared with Multi-task LASSO in our evaluation. However, Multi-task LASSO originally is a method to select a feature set that explains more than one objective variables. In other words, the task we are focusing on, i.e., retrieving a commonly effective feature set among many features under a single objective variable, has not been tackled so far. Nevertheless, this new task may be required in several cases, e.g., analyzing causes of a pandemic disease among countries or regions could be a target. There are many candidate features that causes a disease spreading. To identify the common causes among regions and separate them from those specific to the region may be a useful application.

8 CONCLUSION

In this paper, we proposed a new feature selection method IFS (Incremental Feature Selection) that incrementally increases the number of elements in the candidate feature sets in order to finally select a quasi-optimal feature set that minimizes MSE in multiple regression analysis. In addition, we considered to retrieve commonly effective feature sets among different regions of Wagyu brands to identify the generally effective features that explain Wagyu beef quality regardless of Wagyu regions. We proposed to apply fairness indices among MSE values of multiple regions to limit the least allowable fairness among MSEs as a constraint.

Through evaluation compared with LASSO and Multi-task LASSO, we showed that the proposed method IFS outperforms those conventional methods. From the result, we proved that the incremental approach works more effectively to decrease error in MSE than the well-known LASSO. Although the computational time of IFS is larger than LASSO, it is still feasible if the given feature set size is in the order of hundreds or thousands. Additionally, we showed that IFS enables us to provide a constraint of fairness in MSEs among different Wagyu regions. We outperform MT-LASSO in retrieving a fairly effective feature set not only optimality in minimizing MSEs but also in that it guarantees the minimum fairness in MSE among regions.

As future work, how to determine the parameters J and l would be interesting. In addition, designing a method for predicting beef-quality traits from the common and regional feature sets retrieved from IFS would be one of the challenging topics.

Acknowledgment

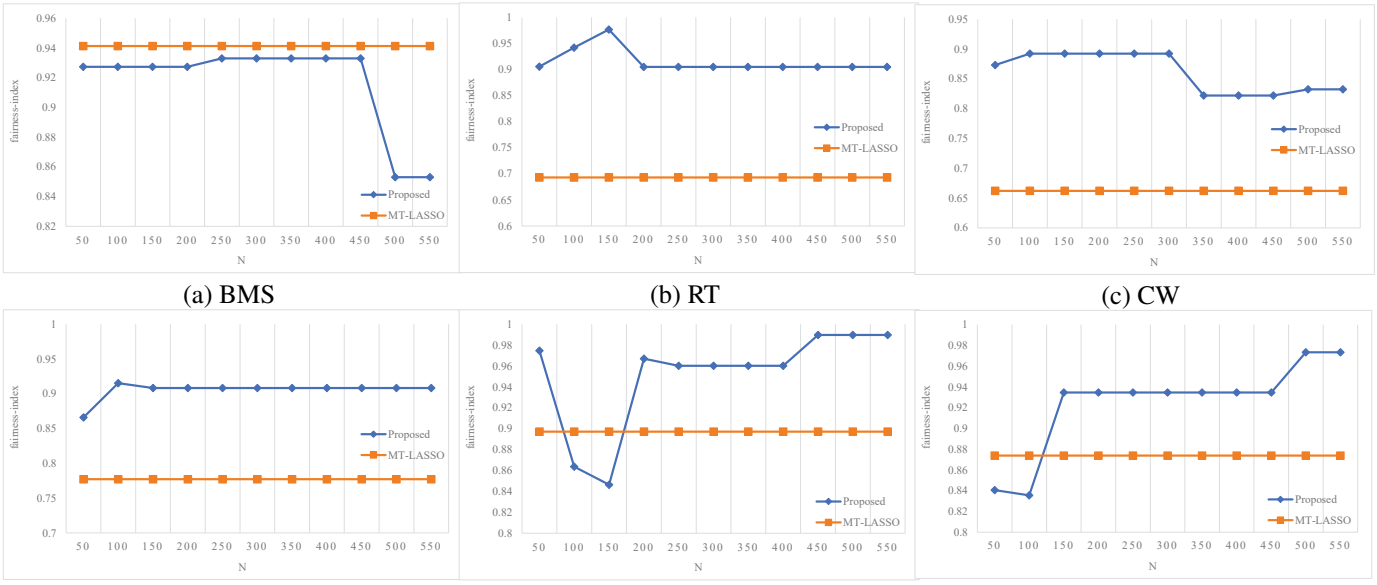
This work is supported by “the Program for Promotion of Stockbreeing” of JRA (Japan Racing Association).

REFERENCES

[1] R. Tibshirani, Regression Shrinkage and Selection via the Lasso, Journal of the Royal Statistical Society, Series B, 58(1), pp.267–288 (1996).

[2] G. Obozinski, B. Taskar, and M. Jordan., Multi-task feature selection, In Technical Report, Department of Statistics, University of California, Berkeley, (2006).

[3] S. Hara and T. Maehara, “Enumerate Lasso Solutions for Feature Selection,” In Proc. AAAI2017, (2017).



(a) BMS (b) RT (c) CW

(d) SFT (e) REA (f) YE

Figure 4: Results on Fairness

[4] R. Jain, D.M. Chiu, W. Hawe, "A Quantitative Measure of Fairness and Discrimination for Resource Allocation in Shared Computer Systems," DEC Research Report TR-301, (1984).
[5] Japan Meat Grading Association, "The Manual of Standard in Dealing Pork and Beef," (2001) (In Japanese).
[6] A Guide of Japanese Black Cattle Sires, http://liaj.lin.gr.jp/index.php/detail/data/m/803237096 (referred in October 2020) (In Japanese).
[7] Wagyu Registry Association,"Compliation of Sires of Japanese Black Cattle," (2003) (In Japanese).
[8] N. D. Cameron, "Selection Indices and Prediction of Genetic Merit in Animal Breeding," CAB International (1997).
[9] A. Oka, T. Dohgo, M. Juen, and T. Saito, "Effects of vitamin A on beef quality, weight gain, and serum concentration of thyroid hormones, insulin-like growth factor-I, and insulin in Japanese black steers," Animal Science and Technology (1998).
[10] C. Ludwig, L. Gillet, G. Rosenberger, S. Amon, B.C. Collins, R. Aebersold, "Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial," Molecular Systems Biology (2018) DOI 10.15252/msb.20178126.
[11] R-Project, https://www.r-project.org/, (referred in October 2020)
[12] Scikit-learn, https://scikit-learn.org/ (referred in October 2020)

(Received May 14, 2020)
(Revised October 28, 2020)

Nanami Higashiguchi received the B.E. degree from Wakayama University in 2019. She is currently a Master-course student in Wakayama University. She is interested in Data Science, Statistical Modeling, and Wa-gyu beef improvement. She is a student menber of IPSJ.

Masatsugu Motohiro received the B.E. and M.E. degree from Wakayama University in 2018 and 2020, respectively, and he is now with Hibiya Computer Systems, Co., Ltd. He is interested in Data Science, Statistical Modeling, Agricultural IoT, and so on.

Haruka Ikegami received her B.E. from Kindai University in 2004. She was a researcher in Kindai University from 2004 to 2020. She is interested in agriculture proteomics, bioinformatics, mass spectrometry, and so on.

Tamako Matsuhashi received her B.A. and M.S. from Hokkaido University in 1997 and 1999, then received her Ph.D. (Doctor of Veterinary) degrees from The University of Tokyo in 2006. She was a chief scientist and then research specialist in Gifu Prefecture from 2006 and 2014, respectively. She is a junior associate professor in Kindai University from 2016. She is interested in reproduction and bioinformatics of livestock in animal science.

Kazuya Matsumoto Received his B.A. degree from Utsunomiya University in 1984, and then M.S. and Ph.D. (Doctor of Agriculture) degrees from Kyoto University in 1986 and 1989, respectively. He was a staff scientist in NT. Science and Tosoh Co. from 1989 to 1995, and visiting scientist in The Institute of Medical Science, The University of Tokyo from 1995 to 1998. In 1998, he moved as an Assistant Professor to Kindai University. He is a Professor in Kindai University. One of his major research themes is applications of bioinformatics analysis for animal science.

Takuya Yoshihiro received his B.E., M.I. and Ph.D. degrees from Kyoto University in 1998, 2000 and 2003, respectively. He was an assistant professor in Wakayama University from 2003 to 2009. He has been an associate professor in Wakayama University from 2009. He is currently interested in the graph theory, distributed algorithms, computer networks, medial applications, and bioinformatics, and so on. He is a member of IEEE, ACM, IEICE, and Senior member of IPSJ.