Background

Notations

We use y_i and $x_{i,j}$ to denote the response and the value of j^{th} descriptive feature for i^{th} observation respectively. y_i and $x_{i,j}$ are also known as response and explanatory variables. Throughout this study, we use the term "feature" and shall not use them interchangebly with "variable", unless necessary, to be consistent with the context of feature selection. We use the upper cases to represent vectors and the boldface bont for matrices. Suppose there are n observations, we write $Y = [Y_1, Y_2, ..., Y_n]^T$ and $X_j = [x_{1,h}, x_{2,j}, ...x_{n,j}]^T$ to denote n-vectors of responses and values of j^{th} features respectively. We also call X_j as a main-effect feature.

Suppose p descriptive features, we define $\mathbf{X} = [x_{i,j}] \in \mathbb{M}^{n \times p}$ as the matrix of descriptive features for all observations. $\mathbf{X}$ can comprise of both continuous and categorical features; however, we shall not distinguish the feature types unless unnecesssary. To represent pairwise interaction between two descriptive features, say k^{th} and j^{th} features, we express

$$X_{k:j} = X_k \times X_j = [x_{1,k} \times x_{1,j}, x_{2,k} \times x_{2,j}, ..., x_{n,k} \times x_{n,j}]^T \text{ for } k \neq j \tag{1}$$

Loosely following notations of generalised linear models [@bilder], we write a linear predictor function or a link function as follow:

$$f(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p = \beta_0 + \sum_{j=1}^{p} \beta_j X_j \tag{2}$$

The β_0, β_1, ... β_p are known as regression coefficients. Note that Equation 2 contains no error term since a link function describes how the mean of Y is related to the linear combination of the explanatory features throught a mathematical function. In a standard linear regression, $f(Y)$ is an identity function:

$$\mathbb{E}(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p \tag{3}$$

$\mathbb{E}(Y)$ is the expectation of Y. Equation 3 is appropriate for a continuous response. Logistic regression model is most commonly used for a binary response i.e. $Y_i \in \{0, 1\}$, The link function of a logistic regression model is $f(\cdot) = \exp(\cdot)/(1 + \exp(\cdot))$. Therefore, $\mathbb{E}(Y)$ can therefore be written as:

$$\mathbb{E}(Y) = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p)} \tag{4}$$

To include pairwise interaction features in a link function, we write:

$$f(Y) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \sum_{j<k} \beta_{j:k} X_{j:k} \tag{5}$$

$\beta_{j:k}$ is the interaction coefficient between k^{th} and j^{th} features. Equation 5 is the basic model formulation in this study.

Aspects of Interactions

@Cox1984 provides a statistical exposition of various aspects of interactions on the structure of $\mathbb{E}(Y)$. He provides a precise definition of "interactions". Given two features X_j and X_k, when a function $f(X_j, X_k)$ cannot be expressed as $h_1(X_k)+h_2(X_j)$ for some functions h_1 and h_2, there exists an interaction in f between X_k and X_k. Cox further classifies $\mathbf{X}$ into *treatment*, *intrinsic*, and *non-specific* features (variables).

The treatment features refer to control factors in observational studies. The intrinsic features are factors beyond the control of experimenters, for examples, age and gender. The non-specific features are known as random factors such as experimental blocks and replicates. By restricting to two-way interactions, Cox argues there are three different kinds to be considered:

- Treatment $\times$ Treatment;
- Treatment $\times$ Intrinsic;
- Treatment $\times$ Non-specific.

In our study, we do not distinguish these kinds above for pragmatic reasons. In practice, machine learning practitioners are no experimenter and hence have no control on features. Also, Cox ignores the case of Intrinsic $\times$ Intrinsic, which can be of interest. For example, there might be interaction between gender and age in estimating an individual's body composition. Mixed effect models can be used to incorporate non-specific factors including replicates and blocks. The mixed effect models are beyond the scope of this study and hence shall not be discussed further.

With a careful formulation, Cox argues that F-tests are powerful to detect interactions of interest whereas graphical methods are appropriate for situations where the interaction cannot be completely estimated. Cox's methodology has a few shortcomings. First, F-tests or other statistical hypothesis tests might be difficult to formulate and graphical methods might be infeasible when there are many features. Second, statisical tests can severely limit the scope of how significant main-effect and interaction features can be selected. For example, models $g(Y) = \beta_0 + \beta_1 X_1 + \beta X_2$ and $g(Y) = \beta_0 + \beta_1 X_1 + \beta X_2 + \beta_{1:2} X_{1:2}$ can be compared using a statistcal hypothesis test to assess if the interaction term of $X_{1:2}$ should be included. However, we cannot *easily* compare models $g(Y) = \beta_0 + \beta_1 X_1 + \beta X_2 + \beta_{1:2} X_{1:2}$ and $g(Y) = \beta_0 + \beta_1 X_4 + \beta X_5 + \beta_{4:5} X_{4:5}$ in a nested framework. Feature selection hence becomes useful to assess any model specification.

When interaction features are included, a typical feature selection process might lead to an outcome where interaction features are selected but their main effects are not. Since it is rarely to have interactions without main effects, @glinternet posit to enforce some degree of hierarchy on model specification. They define a model to follow a strong hierarchy when an interaction is present only if both of its main effects are present too. A model is weakly hierarchical as long as either of the main effect features are present. Anti-hierarchical model is when interactions are only pairs of main effects are absent. A pure-interaction model contains no main-effect features. While the true structure of hierarchy is unknown, anti-hierarchical and pure-interaction form are rare. We develop two-step SP-FSR by assuming a strong hierarchy.