UNIVERSITÉ DE GENÈVE

IOM PhD Day - Friday 18 June 2021

SWAG: A Wrapper Method for Sparse Learning

Stéphane Guerrier

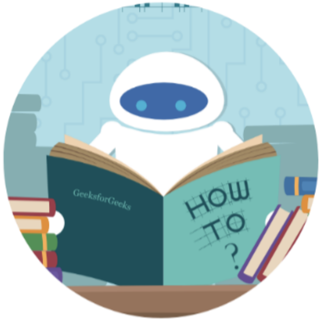
Research Areas

Computational Statistics

Monte Carlo methods
Simulation-based inference

High Dimensional Data

Simulation-based bias correction
Model selection



Machine Learning

Wrapper methods
Applications in Genomics

Signal Processing

Time series analysis
Applications in Robotics

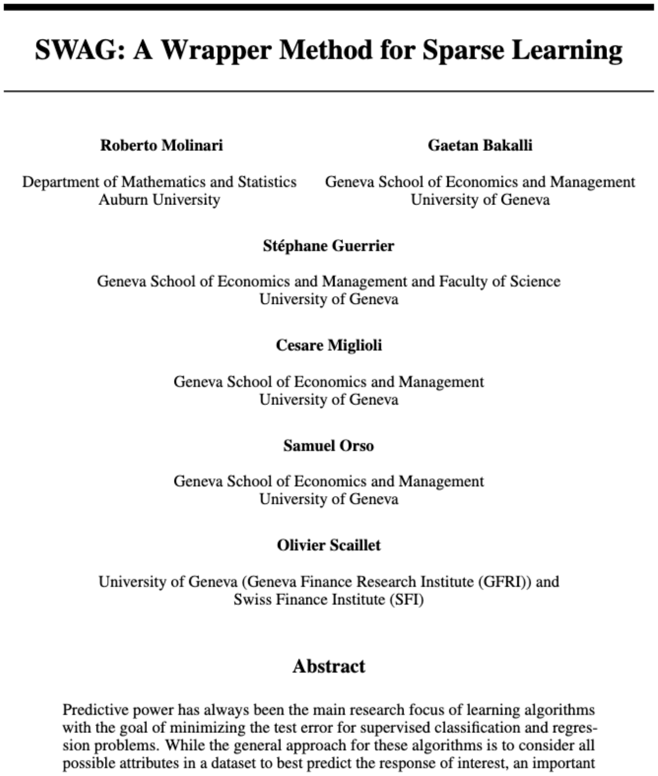
Biostatistics

Bioequivalence
Prevalence estimation

Applied Statistics

Applications in Economics, Biology,
Education and Civil Engineering

A New Wrapper Method for Sparse Learning



SWAG: A Wrapper Method for Sparse Learning

Roberto Molinari

Department of Mathematics and Statistics
Auburn University

Gaetan Bakalli

Geneva School of Economics and Management
University of Geneva

Stéphane Guerrier

Geneva School of Economics and Management and Faculty of Science
University of Geneva

Cesare Miglioli

Geneva School of Economics and Management
University of Geneva

Samuel Orso

Geneva School of Economics and Management
University of Geneva

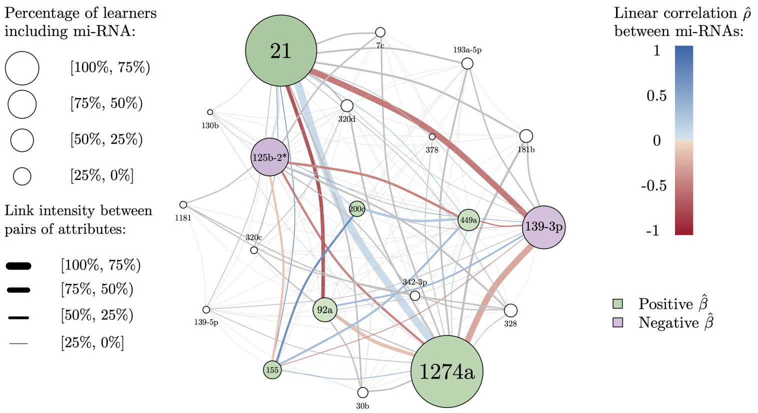
Olivier Scaillet

University of Geneva (Geneva Finance Research Institute (GFRI)) and
Swiss Finance Institute (SFI)

Abstract

Predictive power has always been the main research focus of learning algorithms with the goal of minimizing the test error for supervised classification and regression problems. While the general approach for these algorithms is to consider all possible attributes in a dataset to best predict the response of interest, an important

https://arxiv.org/abs/2006.12837



Example of **predictive network** based on the breast cancer dataset of Haakensen et al. (2016).

SWAG R 📦 available on CRAN here.

Logistic Regression

- In statistics, the **logistic model** is used to model the probability of a certain event existing such as pass/fail, healthy/sick or buying/not buying an item.
- In most (modern) empirical studies, a large number of variables are measured (often larger than the number of observation) and therefore researchers generally try to identify the smallest possible set of variables that can still achieve good predictive performance. This task is often referred to as model (or variable) selection.

How statisticians (typically) understand this definition:

- We are looking for a single model.
- For a given candidate model, picking the most likely parameters given the data is optimal (i.e. maximum likelihood estimation).
- Predictive performance can be measured by the likelihood function (typically out-of-sample criteria such as AIC or BIC).

Is this a good idea? 🤔

According to our understanding of the problem (i.e. single model based on likelihood methods): **YES!**

However, this approach has some drawbacks:

- Focusing on a **single model suggests a level of confidence in our final result that is not justified by the data** as other models generally exist with similar goodness of fit.
- Maximizing the likelihood function does not guarantee finding the best model(s) (and parameters) according to a given out-of-sample objective function.
- These methods are prone to overfitting (due to the asymmetric effects of "under" vs "over" fitting).

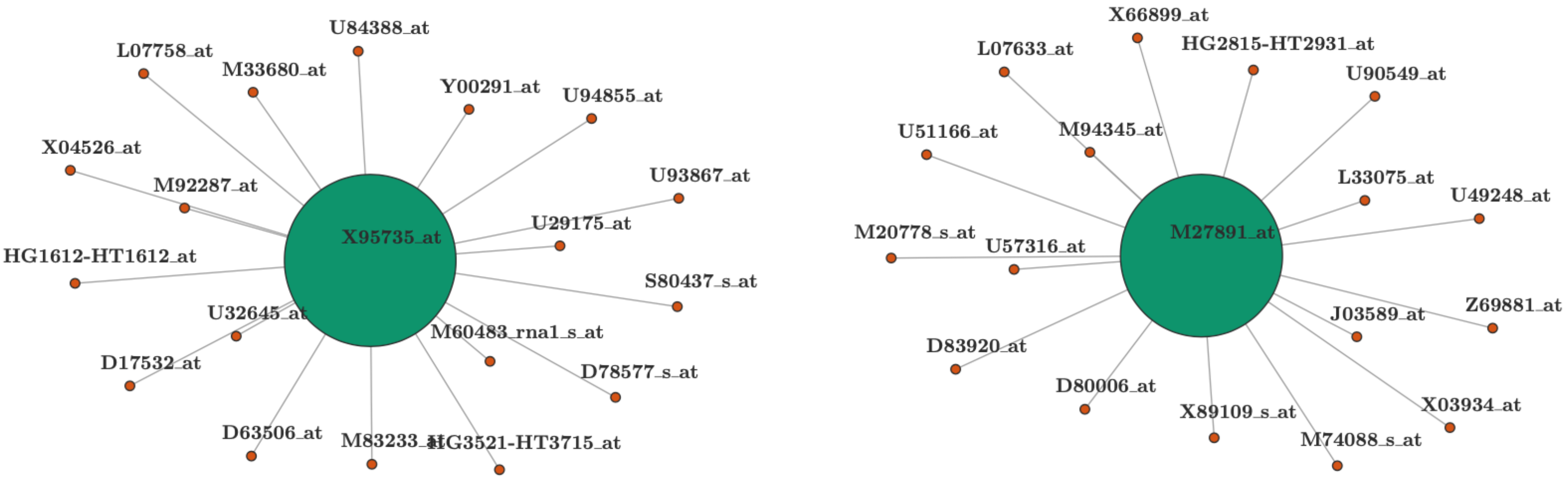
How to address these limitations?

- A possible solution is the SWAG algorithm proposed in Guerrier et al. (2016) and later improved in Molinari et al. (2021). This wrapper method aims to alleviate (some of) these issues and is built "on top" of an arbitrary predictive method or algorithm (e.g. logistic regression, random forest, ANN, ...).
- Its goals are the following:
  - Finding ALL models (and parameters) minimizing an out-of-sample user-defined objective function (e.g. classification error).
  - Restricting our attention to the models with the smallest dimension.
- As an example, we applied in Guerrier et al. (2016) our methodology to a well-known dataset on the classification of Acute Myeloid Leukemia (AML) against Acute Lymphoblastic Leukemia (ALL) (see Golub et al., 1999).

Leukemia Dataset

Classification error

SWAG-based network

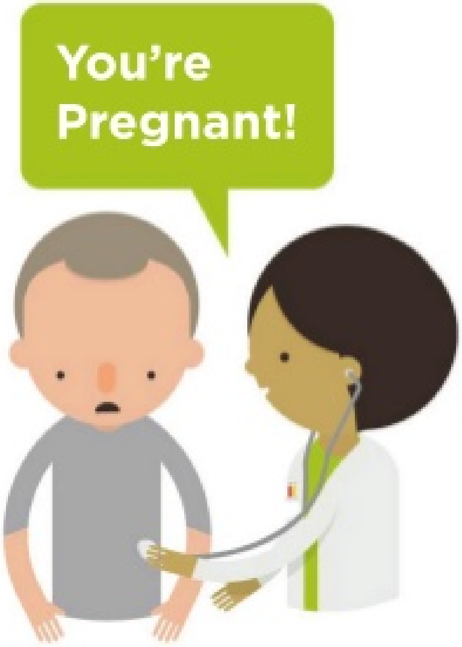


Adapted from Guerrier et al. (2016).

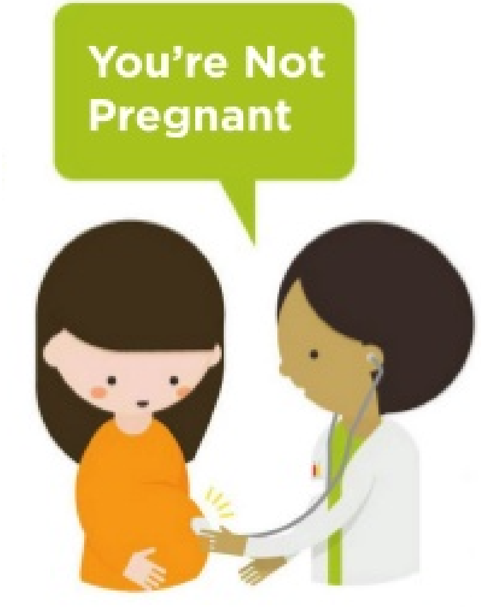
Context of the analysis

- Most statistical methods are used to deliver estimates and predictions for different problems with the goal of using these results for decision-making (e.g. whether a patient should be treated, selecting a drug or its dosage, marketing strategy, ...).
- The vast majority of these methods are **independent** of the context of this decision-making process and the approach used to obtain model's parameters (and prediction) is **fixed** for all applications.
- This can lead to contradictory examples (e.g. asymmetric "risk" of treating vs not treating a patient). Indeed, not all prediction errors are the "same" or have similar "impacts" (e.g. missing on a good investment opportunity or investing in a bad one).
- The SWAG allows to incorporate the "context" of the decision-making through its user-defined objective function (e.g. weighted risks).

Not all classification errors are the same... 🤔



Type I error (false positive)



Type II error (false negative)

Example: Customer churn

- Customer churn or attrition corresponds to the loss of clients or customers.
- Banks, telephone service companies, internet service providers and other similar companies often use customer attrition rates as one of their key business metrics.
- Adequate (probabilistic) modeling of customer churn is therefore a matter of interest. A classical and standard approach to model such data is following:
    - Model the data using a logistic regression.
    - Select the variables using the AIC (or other similar criteria) in a stepwise (greedy) manner (a popular alternative would be based on the lasso).
    - Make inference on the variables of interest based on the MLE of the selected model (⚠ as if the model was known! ⚠).

Customer churn: dataset

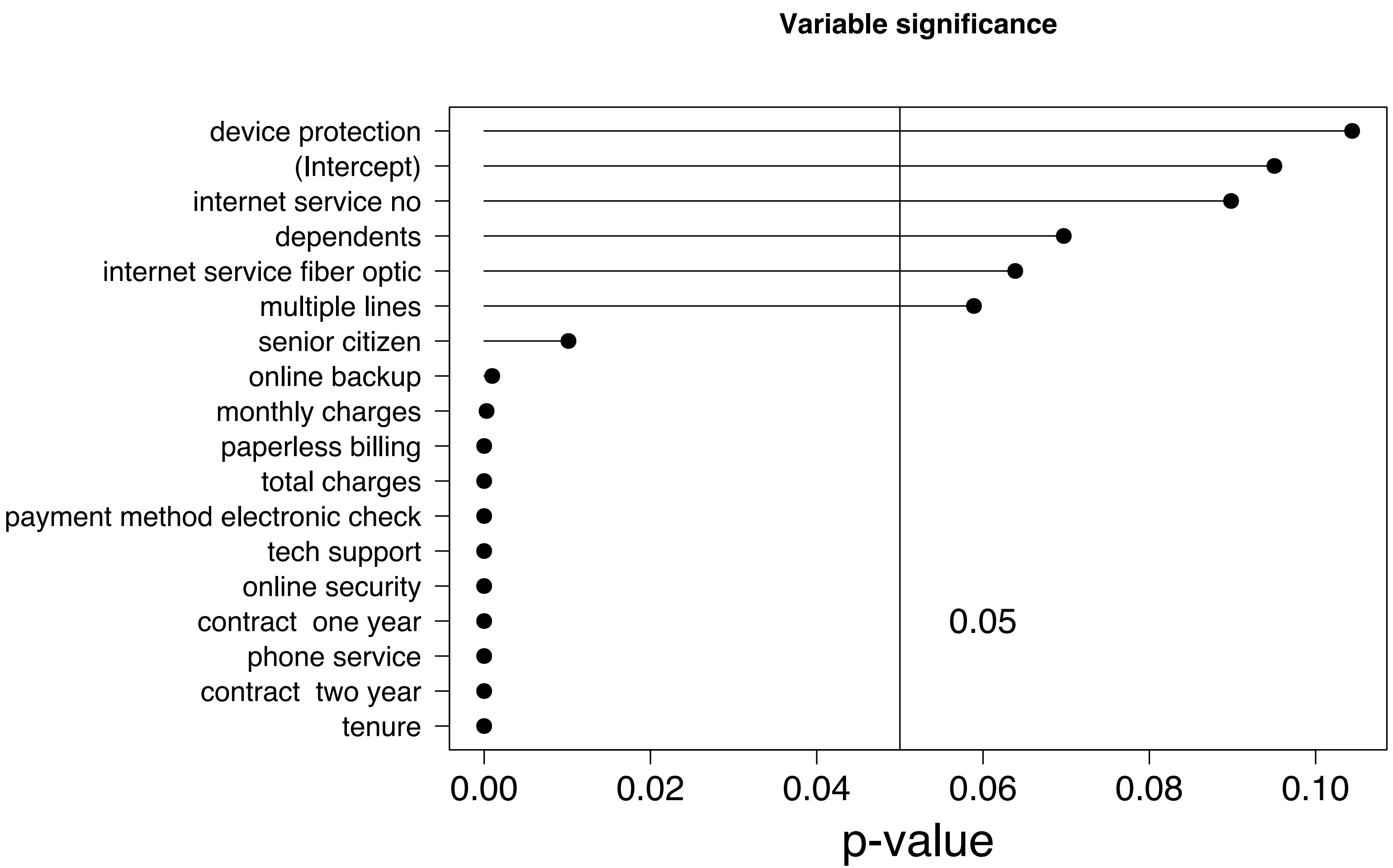
We consider the *Telco customer churn* that provides information about customers of a telecom company and whether or not they left the company.

The dataset includes information about:

- Customers who left within the last month.
- Services that each customer has signed up for (e.g. phone, multiple lines, internet, online security, ...).
- Customer account information (e.g. how long they have been a customer, contract, payment method, ...).
- Demographic info (e.g. gender, age range, ...).

This dataset is relatively large with 10^4 observations and more than 50 variables.

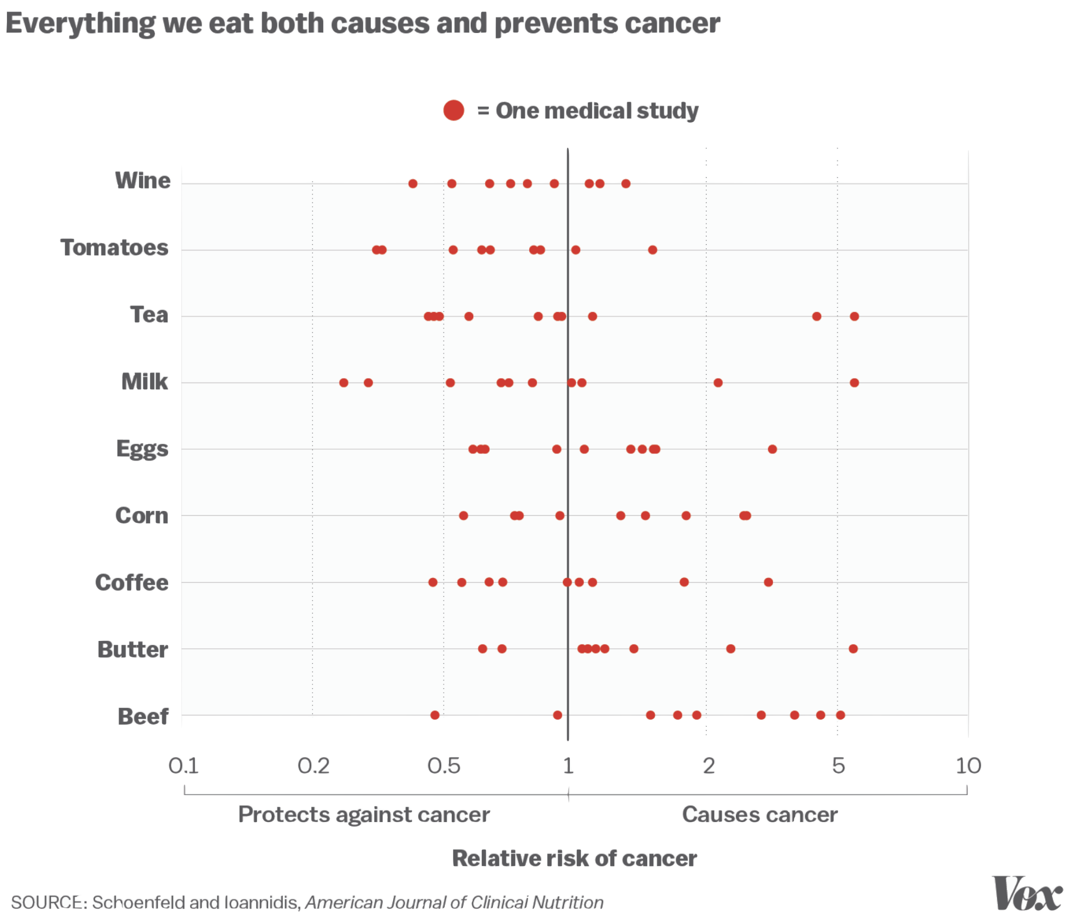
Customer churn: standard analysis



Customer churn: SWAG-based analysis

Scree plot Density Network

Thank you very much for your attention!



👋 Read the original article: "*This is why you shouldn't believe that exciting new medical study*" here.