Goodness-of-fit Testing in High Dimensional Generalized Linear Models - Using Toxicity Dataset

Sandamini Senaratne

November 28, 2023

Overview

1. Journal article used
2. Target problem
3. Data Description
4. Statistical Problem
5. Suggested Solution
6. Model and Estimation used
7. Methods Used
8. Results
9. Discussion
10. References

Journal article used

Goodness-of-fit testing in high-dimensional generalized linear models. Jana Janková, Rajen D. Shah, Peter Bühlmann, Richard J. Samworth

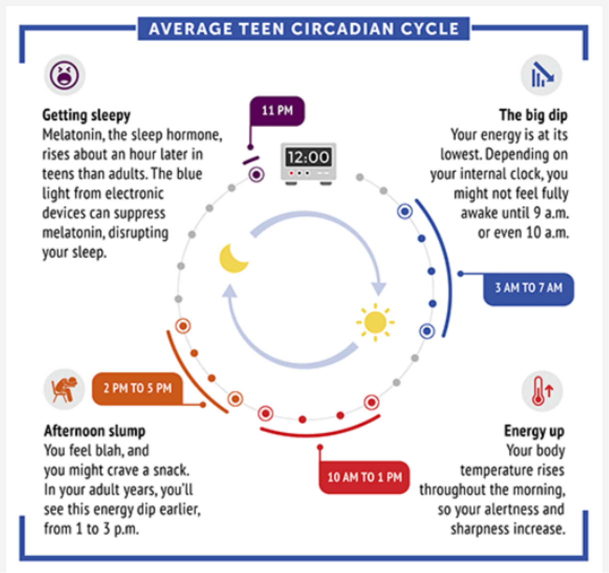
- A family of tests to assess the goodness-of-fit of a high dimensional generalized linear model are proposed.
- A new method for detecting conditional mean mis-specification in Generalized Linear Models on the basis of predicting remaining signals in the residuals is introduced.

Target problem

Assessing the goodness-of-fit of the Generalized Linear Model which consists of molecular descriptors responsible for generating circadian rhythms.

- Circadian rhythms are physical, mental, and behavioral changes that follow a 24-hour cycle
- Example: Sleeping at night and being awake during the day

Circadian rhythm cycle of a typical teenager



AVERAGE TEEN CIRCADIAN CYCLE

Getting sleepy
Melatonin, the sleep hormone, rises about an hour later in teens than adults. The blue light from electronic devices can suppress melatonin, disrupting your sleep.

The big dip
Your energy is at its lowest. Depending on your internal clock, you might not feel fully awake until 9 a.m. or even 10 a.m.

Afternoon slump
You feel blah, and you might crave a snack. In your adult years, you'll see this energy dip earlier, from 1 to 3 p.m.

Energy up
Your body temperature rises throughout the morning, so your alertness and sharpness increase.

Data Description

Dataset have been taken from the UCI Machine Learning Repository: Toxicity Dataset

Observations include 171 molecules designed for functional domains of a core clock protein, CRY1, responsible for generating circadian rhythms

- CRY1, the circadian cryptochrome, is a pro-tumorigenic factor that rhythmically modulates DNA repair

Consists of 1203 variables: a complete set of molecular descriptors

The dependent variable is a binary variable which classifies whether the molecule is toxic or non-toxic

Statistical Problem

The goodness-of-fit test methods for Generalized Linear Models rely on the properties that hold only on low dimensional setting such as asymptotic linearity and normality of the likelihood estimator.

These test methods may fail when there is large number of covariates in the model.

Focus on the detection of misspecification in the fit of a high dimensional Generalized Linear Model.

Suggested Solution

- First fit a lasso-penalized Generelized Linear Model to get the β estimates

$$\hat{\beta} := \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^{n} \rho(Y_i, x_i^T \beta) + \lambda \|\beta\|_1 \right\}$$

- Use the β estimates to predict the Pearson type residuals using

$$R_i = \frac{Y_i - \mu(x_i^T \hat{\beta})}{\sqrt{V\{\mu(x_i^T \tilde{\beta})\}}}$$

Model and Estimation for goodness-of-fit testing on logistic regression model

We take $\mathcal{Y} = \{0, 1\}$ and assume that $(Y_i | x_i = x)$ $Bernoulli\{\pi_0(x)\}$. Define

$$f_0(x) := log\left\{\frac{\pi_0(x)}{1 - \pi_0(x)}\right\}$$

i.e $\mathbb{E}(Y_i | x_i = x) = \pi_0(x) = \mu\{f_0(x)\}$, for the link function $\mu(u) = 1/\{1 + exp(-u)\}$

The l_1 -regularized logistic regression estimator is

$$\hat{\beta} := \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \left\{ -Y_i {x_i}^T \beta + d({x_i}^T \beta) + \lambda \parallel \beta \parallel_1 \right\}$$

where $d(\xi) := log\{1 + exp(\xi)\}$

Methods Used

Logistic regression was fitted using several types of methods and obtained a parameter estimate $\hat{\beta}$ with its corresponding support set $\hat{S}$. Then the GRP test was conducted using the respective outputs.

- Lasso
- Ridge
- Adaptive Lasso
- Debiased Lasso

Results

Three different forms for the misspecification g(·) were considered:
$g(u) = 0, \quad g(u) = u_{j1}^2 + u_{j2}^2, \quad g(u) = u_{j1}u_{j2} + u_{j3}u_{j4}$

The rejection probabilities for all three scenarios from 100 repetitions were reported.

	lasso	ridge	adaptive lasso	de-biased lasso
$g(u) = 0$	0.11	0.07	0.12	0.09
$g(u) = u_{j1}^2 + u_{j2}^2$	0.39	0.46	0.36	0
$g(u) = u_{j1}u_{j2} + u_{j3}u_{j4}$	0.55	0.11	0.09	0

Discussion

In each case, the GRP-test is able to detect the misspecification relatively reliably, while keeping the type I error under control.

De-biased lasso method might not be effective in capturing the linear component of the model.

The parameters used or the data itself might be leading to degenerate solutions, causing the algorithm to terminate without finding a non-zero solution.

References

Jana Jankov´a, Rajen D. Shah, Peter B¨uhlmann and Richard J. Samworth (2020)
Goodness-of-fit testing in high-dimensional generalized linear models
Journal of the Royal Statistical Society 82(3), 773-795.

Seref Gul, F. Rahim, Safak Isin, Fatma Yilmaz, Nuri Ozturk, M. Turkay, I. Kavakli (2021)
Structure-based design and classifications of small molecules regulating the circadian rhythm period

https://github.com/jankova/GRPtests/tree/master

https://www.stat.math.ethz.ch/ geer/Atlanta3.pdf

https://arxiv.org/pdf/1408.4026.pdf

https://pubmed.ncbi.nlm.nih.gov/14693814/

Thank You!