Spring 2013

Group testing models with unknown link function

Dewei Wang

Karunarathna B. Kulasekera

Colin M. Gallagher

Christopher S. McMahan

Group testing models with unknown link function

Dewei Wang[1)*], K. B. Kulasekera[2)], Colin Gallagher[3)], and Christopher McMahan[4)]
1, 3, 4 : Department of Mathematical Sciences, Clemson University
2 : Department of Bioinformatics and Biostatistics, University of Louisville

1 Introduction

Let us consider the problem of screening a **large number** of individuals for an infectious disease. Traditionally, a specimen (e.g., blood, urine, plasma, etc.) is collected from each of the individuals and is subsequently tested for the presence of the infection:

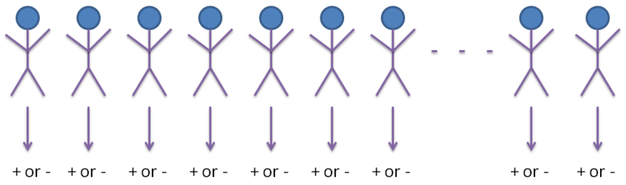

Figure 1. Individual testing.

Due to the large number of individuals, this process can be both expensive (with respect to testing cost) and time consuming.

Group testing, also known as pooled testing, was first proposed by Dorfman in 1943 as a method for reducing the cost associated with screening World War II soldiers for syphilis. In general, group testing involves testing pooled specimens formed from amalgamating specimens collected from individuals, rather than testing each specimen separately:

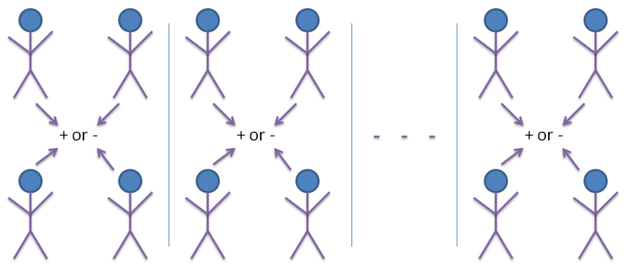

Figure 2. Group testing with size four.

Information that we gain from this pooled testing process are:

- If a group tests negative, then we may conclude that all contributing individuals are negative.
- If a group tests positive, then we may conclude that there is at least one positive individual in the pool.

When testing for low prevalence diseases, pooling specimens has become a common method of increasing screening efficiency. In practice, group testing strategies have been successfully applied in a variety of areas, including genetics, bioterrorism detection, and drug discovery.

Statistical research in group testing are branched into two major area:

- **Classification**: How to design an efficient decoding algorithm so that one can diagnose all individuals as either positive or negative with minimum cost. The basic idea here is to retest individuals' specimen in positive pools.
- **Estimation**: How to estimate individuals' disease risks (the probability of being infected) by use pool response data only. **In this poster, I will present a new methodology which can flexibly model this probability while maintaining a good interpretability.**

2 Notation

- We consider the situation in which N individuals are to be screened for a binary characteristic of interest.
- Further, we assume each of these individuals are assigned to a pool of size n_j, for $j = 1, ..., J$.
- Let Y_{ij} denote the true (latent) status of the ith individual in the jth pool, such that

$$Y_{ij} = \begin{cases} 1, \text{ truly positive,} \\ 0, \text{ truly negative.} \end{cases}$$

- For modeling, we assume that $\boldsymbol{X}_{ij}$, a p-dimensional vector of predictor variables is available for each individual and

$$P(Y = 1 | \boldsymbol{X} = \boldsymbol{x}) = p(\boldsymbol{x}^\top \boldsymbol{\beta})$$

where $p(\cdot)$ is an **unknown link** function. **The major contribution** in our methodology is that we allow the link function to be unknown and then estimate it from the data.

- Since tests are carried on pools instead of on individuals, we cannot observe Y_{ij}s. The only information we have are the testing responses for pools; i.e.,

$$Y_j^* = \max\{Y_{1j}, \ldots, Y_{n_j j}\}.$$

If $Y_j^* = 0$, all the Y_{ij}, $i = 1, \ldots, n_j$, are zero. If $Y_j^* = 1$, at least one of them are one, but the problem is that we do not know which of them are one.

3 Methodology

The methodology we investigated could be summarized as in the following figure. We first estimate the dependence of the link function $p(\cdot)$ on the coefficient β. To emphasize this dependence, we denote it as $p_{\boldsymbol{\beta}}(\cdot)$. Then we maximize the estimated likelihood of the data with respect to β to obtain the final estimator of β.

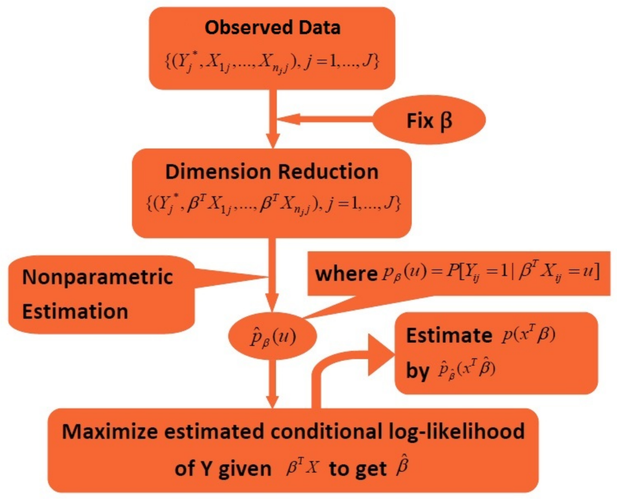

Figure 3. The procedure of our proposed method.

4 Numerical Analysis

We consider the following models:

(1) : $p(\boldsymbol{x}^\top\boldsymbol{\beta}) = \exp(-1 + 5\boldsymbol{x}^\top\boldsymbol{\beta})/[8 + 8\exp(-1 + 5\boldsymbol{x}^\top\boldsymbol{\beta})]$
(2) : $p(\boldsymbol{x}^\top\boldsymbol{\beta}) = \exp(\boldsymbol{x}^\top\boldsymbol{\beta})/20$
(3) : $p(\boldsymbol{x}^\top\boldsymbol{\beta}) = (\boldsymbol{x}^\top\boldsymbol{\beta})^2/8$
(4) : $p(\boldsymbol{x}^\top\boldsymbol{\beta}) = [\sin(\pi\boldsymbol{x}^\top\boldsymbol{\beta}) + 1.2]/[20 + 160(\boldsymbol{x}^\top\boldsymbol{\beta})^2\{sign(\boldsymbol{x}^\top\boldsymbol{\beta}) + 1\}]$

Simulation settings:

- $\boldsymbol{\beta} = (\sqrt{3}, 1)^\top/2 = (0.866, 0.500)^\top$
- $\boldsymbol{X}$: $X_1 \sim U[-1, 1]$, $P(X_2 = \pm 0.5) = 0.5$
- $N = 5000$, $n_j = n \in \{1, 2, 5, 10\}$

The above data generating process was repeated 200 times, for each model and setting, and our methodology was applied to each. The following figure and table summarize the behavior of the 200 resulting estimates.

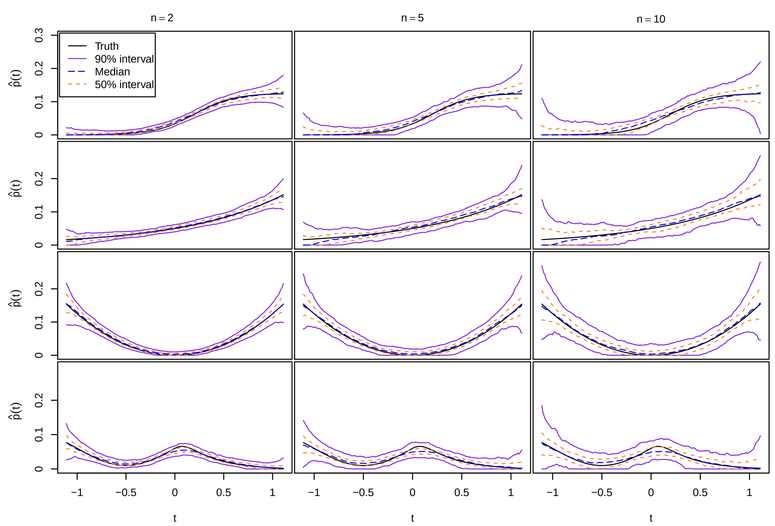

Figure 4. Quantile curves of the 200 estimates of the link function $p(\cdot)$

One can interpret the 90% interval curves as a "confidence band" constructed by our estimate. And you can see that the truth is well covered by this band and the width of this band is not large. It indicates that our estimate of the link function behaves quite well.

The main message from the following table is that the regression coefficient β can be well estimated by our procedure. One may observe that the standard deviation increases as group size becomes larger. It is natural since that more information are hidden in a larger pool. The same pattern can be found in Figures 4–5.

Table 1. Mean (standard deviation) of the 200 estimates of the coefficient $\boldsymbol{\beta}$

Model	*Target*	$n_j = 1$	$n_j = 2$	$n_j = 5$	$n_j = 10$
(1)	$\beta_1 = 0.866$	0.868(0.032)	0.863(0.042)	0.864(0.063)	0.866(0.102)
	$\beta_2 = 0.500$	0.492(0.058)	0.497(0.076)	0.484(0.122)	0.448(0.196)
(2)	$\beta_1 = 0.866$	0.862(0.058)	0.850(0.084)	0.826(0.128)	0.829(0.146)
	$\beta_2 = 0.500$	0.492(0.105)	0.500(0.142)	0.509(0.209)	0.487(0.234)
(3)	$\beta_1 = 0.866$	0.866(0.024)	0.860(0.042)	0.843(0.079)	0.836(0.097)
	$\beta_2 = 0.500$	0.497(0.041)	0.504(0.069)	0.519(0.116)	0.518(0.157)
(4)	$\beta_1 = 0.866$	0.865(0.028)	0.874(0.037)	0.865(0.086)	0.862(0.107)
	$\beta_2 = 0.500$	0.499(0.048)	0.480(0.069)	0.474(0.137)	0.464(0.174)

5 Hepatitis B data

We analyzed a real hepatitis B data set from NHANES 2009-2010. The is data set consists of 6533 individual observations. Each observation has six variables:

- Y is binary, indicating the presence ($Y = 1$) or absence ($Y = 0$) of the antibodies to the hepatitis B core antigen in the patient's serum or plasma.
- X_1 is age (continuous).
- X_2 is gender (discrete).
- X_3 is the cholesterol level (continuous).
- X_4 is the alanine aminotransferase level (continuous).
- X_5 is ethnicity (discrete).

We considered group sizes $n_j = n = 1, 2, \ldots, 10$. After randomly grouping individuals, we artificially generate the pooled testing response by

$$Y_j^* = \max\{Y_{1j}, \ldots, Y_{nj}\}.$$

Then estimates are computed by our methodology. This procedure is repeated 200 times. The pattern of these 200 estimates are summarized in the following figure.

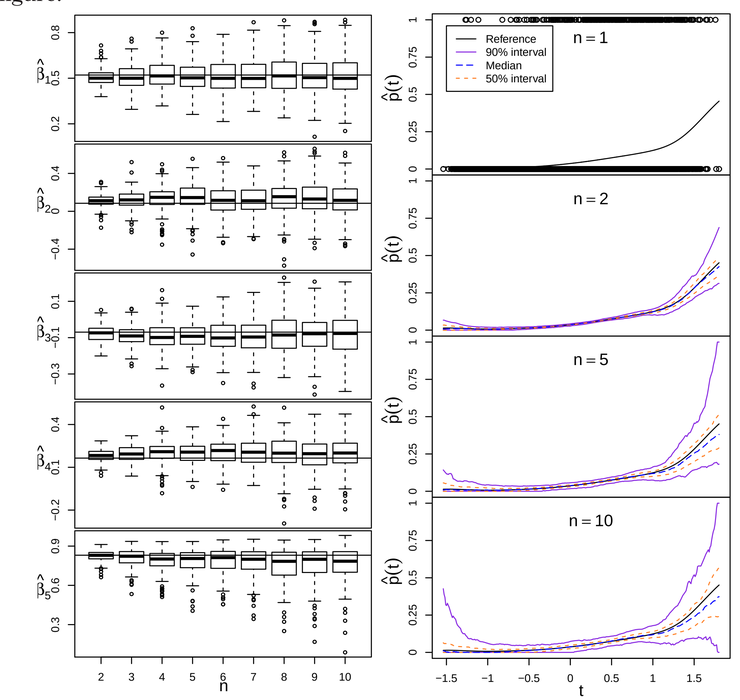

Figure 5. Boxplot of estimates of $\boldsymbol{\beta}$ and Quantile curves of estimates of the link function $p(\cdot)$

6 Conclusion

We have proposed a new method for modeling data collected form a group testing scheme which has becomes a standard procedure for screening a large number of individuals for infectious diseases. Numerical investigation and a real data analysis have demonstrated the performance of our estimators under practical settings. We also extend our method to cover the cases of imperfect testing and missing covariate information. If you are interested in this work, you are very welcome to contact me at *dwang@clemson.edu*.