PBHL-B574 (2023) Methodology-Focused Final Project:

Penalized GEE for High-Dimensional Longitudinal Data Analysis

Yao Chen

May 2, 2023

1. Introduction

Longitudinal studies are research designs that collects data at multiple time points from the same individuals, or experiment units. The goal of longitudinal studies is to characterize the patterns of change in response over time and investigate the factors that contribute to these changes. Analyzing longitudinal data poses several challenges, especially when dealing with a large number of covariates, in other words, high-dimensional data. This type of longitudinal data has become prevalent in large-scale long-term health studies and gene expression experiments. However, in many cases, even with a large number of covariates, only a subset of them is important for accurately and efficiently modeling the response variable. Therefore, to address the challenges of variable selection and estimation in high-dimensional longitudinal data, Wang et al. (2012) proposed the penalized generalized estimating equations (GEE) procedure and this final project of PBHL-B574 is designed based on this methodology paper.

Liang and Zeger (1986) developed longitudinal data analysis using generalized linear models (GLMs) based on the concept of the generalized estimating equations (GEE). The GEE approach has been widely applied in longitudinal studies and provides a consistent estimator, even with misspecified working correlation structure. It is a powerful tool for analyzing correlated data from longitudinal studies or clustered data, however in the

context of high-dimensional data, GEE faces challenges with variable selection and parameter estimation. The large number of covariates compared to the number of observations can lead to overfitting, where GEE models may incorporate numerous irrelevant covariates. Consequently, this can adversely affect generalization and result in low prediction accuracy.

The penalized GEE procedure solves generalized estimating equations with a non-convex penalty function. Similar to GEE, the penalized GEE only requires specifying the first two marginal moments and a working correlation matrix. The addition of penalty functions in penalized GEE helps address the high-dimensional problem by shrinking the coefficients of irrelevant covariates, leading to more accurate and efficient results. Inan and Wang (2017) also introduced an R package PGEE for the implementation of the penalized generalized estimating equations (GEE) procedure. The asymptotic theory of the penalized GEE in a high-dimensional framework was discussed by Wang in 2012, however would not be proved in detail in this report.

This report focuses on reviewing the penalized GEE, replication of the simulation studies with correlated normal responses using its R package `PGEE`.

2. Methods

2.1 Generalized estimating equations

For each subject i measured at jth time points, we have a response variable $\mathrm{Y_{ij}}$ with $\mathbf{Y_i} = \left(Y_{i1},\ \ldots,\ Y_{im_i}\right)^{\mathbf{T}}$ and a p_n-dimensional vector of covariates X_{ij} with $\mathbf{X_i} = \left(\mathbf{X_{i1}}, \ldots,\ \mathbf{X_{im_i}}\right)^{\mathbf{T}}$, $i = 1, \ldots, n,\ j = 1, \ldots, m_i$. Let $\mathbf{A_i}(\beta_n) = diag\left(\sigma_{i1}^2(\beta_n),\ \ldots, \sigma_{im}^2(\beta_n)\right)^T$, a diagonal matrix with variance of Y_{ij} as the jth element, and let $\mathrm{R}(\tau)$ be the correlation

matrix of Y_i, then the variance of Y_i can be written as $\mathrm{V_i} = \mathrm{A_i^{\frac{1}{2}}}(\beta_n)\ \mathrm{R}(\tau)\mathrm{A_i^{\frac{1}{2}}}(\beta_n)$. Assume marginal density of Y_{ij} comes from a canonical exponential family, the estimating equations can be reduced to:

$$n^{-1}\sum_{i=1}^{n} \mathbf{X_i^T}\ \mathbf{A_i^{\frac{1}{2}}}(\boldsymbol{\beta_n})\ \widehat{\mathbf{R}}^{\mathbf{-1}}\ \mathbf{A_i^{-\frac{1}{2}}}(\boldsymbol{\beta_n})\ (\mathbf{Y_i} - \boldsymbol{\mu_i}(\boldsymbol{\beta_n}))\ = 0$$

The GEE estimates the population-average or marginal effect of the predictors on the outcome variable, rather than the subject-specific effect. To account for the correlation of within-subject data, GEE specifies a working correlation matrix structure $\mathrm{R}(\tau)$, such as independence, AR(1), or exchangeable. The Huber-White "sandwich estimator" provides robustness when the working correlation matrix structure is misspecified. Unlike likelihood-based methods, GEE is a quasi-likelihood method, which means only the first two moments, the mean and the covariance matter. However, there is no readily available goodness-of-fit measure for GEE, making model selection unclear.

2.2 Penalized generalized estimating equations

Let $S_n(\beta_n)$ be the left-hand side of the estimating equation in GEE, $q_{\lambda_n}(|\beta_n|)$ be a p_n-dimensional vector of penalty functions. The penalized generalized estimating functions are defined as

$$U_n(\beta_n) = S_n(\beta_n) - q_{\lambda_n}(|\beta_n|)sign(\beta_n)$$

where $sign(t) = I(t > 0) - I(t < 0)$. The turning parameter λ_n determines the amount of shrinkage.

The non-convex SCAD penalty function achieves three desirable properties of variable selection: unbiasedness, sparsity and continuity, while LASSO (L_1 penalty)

does not satisfy the unbiasedness condition. The paper advised to continue with SCAD penalty function.

2.3 Simulations

We evaluate performance of variable selection and model fitting of traditional GEE, penalized GEE, and the oracle GEE with prior knowledge of the true marginal regression model for normal responses. For three models, we consider three different working correlation structures, independence, exchangeable, and AR(1). We use a fourth-fold cross validation to select the tuning parameter λ_n. For each simulation, we generate 100 data sets, and run 30 iterations for each data set with the iterative algorithm to estimate β_{n0}. The estimation accuracy and model selection performance is evaluated by:

(a) Mean square error (MSE) between the estimated beta, and the true beta values.

(b) The proportion of times the methods under-selecting (U), over-selecting (O) and exactly selecting (EXACT) the number of relevant covariates in the model. An estimated coefficient of less than 10^{-3} is considered as zero.

(c) The average false positives (FP), the average count of selected non-zero covariates that correspond to the zero coefficients in the underlying model, and the average true positives (TP), the average count of selected non-zero covariates that correspond to the non-zero coefficients in the underlying model.

The correlated normal responses are generated from $Y_{ij} = X_{ij}^T\beta + \epsilon_{ij}$, with 200 subjects, 4 timepoints, and 200 covariates. We generate X_{ij} from the multivariate normal distribution with mean 0 and an AR(1) covariance matrix with marginal variance

1 and auto-correlation coefficient 0.5. The random errors are also generated from the multivariate normal distribution with marginal mean 0, marginal variance 1 and an exchangeable correlation matrix with parameter ρ. We consider $\rho = 0.5$ and 0.8 to represent different strength of within cluster correlation.

The replicated simulation study uses R package `PGEE` on CRAN. The package contains three functions, `CVfit`, `PGEE`, and `MGEE`, for computing tuning parameter, fitting penalized GEE for longitudinal data with high-dimensional data, and fitting unpenalized GEE, correspondingly.

3. Results

Table 1 in appendix summarizes the estimation accuracy and model selection properties of the penalized GEE, the unpenalized GEE and the oracle GEE for three different working correlation matrices and two different values of strength of within cluster correlation ρ. The oracle GEE with the true parameters known in advance shows the best estimation accuracy as it is expected to, and the penalized GEE procedure performs closely to the oracle GEE with significantly lower MSE than unpenalized GEE estimator. Using the true correlation structure (exchangeable) in penalized GEE gives the smallest MSE. Furthermore, it can be seen that the unpenalized GEE generally selects redundant covariates. The penalized GEE successfully selects all covariates with nonzero coefficients (TP=4) and has a small number of false positives, mostly less than 4.

Results of our simulation study are shown in Table 2 in appendix. Similar to the original paper, the oracle GEE with the true parameters known as initial β values show

the best estimation accuracy, and the penalized GEE procedure performs closely to the oracle GEE with significantly lower MSE than unpenalized GEE estimator. However, these MSEs are suspicious because using the true correlation structure (exchangeable) does not provide the smallest MSE. We have investigated into this with different initial β values and concluded the problem may be coding with the R package `PGEE`.

The overfitting trend of unpenalized GEE still exists (FP>190). The penalized GEE with independence working correlation matrices successfully selects all covariates with nonzero coefficients (TP=4) and has a fairly small number of false positives, less than 3. However, the penalized GEE with exchangeable and ar1 working correlation matrices appears an underfitting trend.

When we repeat above steps with only 5 data set generated for each simulation and change initial β values in penalized GEE and unpenalized GEE from zero to null, oracle GEE with known true parameters provides same MSE as unpenalized GEE, which is not expected. The values of MSE are more reasonable comparing with Table 2. The tendency of under-selecting in penalized GEE no longer exists. The penalized GEE successfully selects all covariates with nonzero coefficients (TP=4) and has a small number of false positives.

4. Discussions

Penalized Generalized Estimating Equations (GEEs) offer several advantages in the analysis of high-dimensional longitudinal data. They are flexible, as only specifying the first two marginal moments and a working correlation structure is required. The consistency of model selection is valid, even when the working correlation structure is

misspecified. Furthermore, the paper by Wang (2012) has shown the asymptotic properties of penalized GEEs in a high-dimensional framework where the number of covariates p_n increases as the number of clusters n increases, and p_n can reach the same order as n. In practice, the performance of penalized GEEs has been demonstrated in various applications. Huang et al. (2022) has applied penalized estimating equations and made publicly available the code necessary to reproduce the results.

While we applied the `PGEE` package on CRAN, we noticed that Huang (2022) used penalized GEE with original codes for penalized GEE. Also, our code with `PGEE` package produces suspicious results. We may also improve the simulation results by writing original R codes. The computational complexity of simulation studies has limited the number of data set for each simulation to 100. Penalized GEE would be increasingly popular and easy to validate if the efficiency of computing could be improved. Though penalized GEE performs well, overfitting still is a problem, and its practical performance could be improved with the choice of the tuning parameter.

References

Huang, Y., & Pan, J. (2022). Penalized joint generalized estimating equations for longitudinal binary data. *Biometrical journal. Biometrische Zeitschrift*, *64*(1), 57–73. https://doi.org/10.1002/bimj.202000336

Inan, G., & Wang, L. (2017). PGEE: An R package for analysis of longitudinal data with high-dimensional covariates. *R Journal*, *9*(1), 393-402. https://doi.org/10.32614/rj-2017-030

Wang, L., Zhou, J., & Qu, A. (2012). Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics*, *68*(2), 353–360. https://doi.org/10.1111/j.1541-0420.2011.01678.x

Zeger, S. L., & Liang, K.-Y. (1986). Longitudinal Data Analysis for Discrete and Continuous Outcomes. *Biometrics*, *42*(1), 121–130. https://doi.org/10.2307/2531248

Appendix

Table 1

Original paper: correlated continuous reponse ($n = 200, p_n = 200$) with three different working correlation structures (independence, exchangeable, and AR(1)) and two different strengths of within cluster correlation

	MSE	U	O	EXACT	TP	FP
$\boldsymbol{\rho = 0.5}$						
GEE.indep	0.568	0.00	1.00	0.00	4.00	193.02
GEE.exch	0.381	0.00	1.00	0.00	4.00	192.45
GEE.ar1	0.458	0.00	1.00	0.00	4.00	192.66
oracle.indep	0.009	-	-	-	-	-
oracle.exch	0.006	-	-	-	-	-
oracle.ar1	0.007	-	-	-	-	-
PGEE.indep	0.009	0.00	0.85	0.15	4.00	2.02
PGEE.exch	0.008	0.00	0.33	0.67	4.00	3.30
PGEE.ar1	0.008	0.00	0.38	0.62	4.00	3.00
$\boldsymbol{\rho = 0.8}$						
GEE.indep	0.568	0.00	1.00	0.00	4.00	193.01
GEE.exch	0.165	0.00	1.00	0.00	4.00	190.44
GEE.ar1	0.211	0.00	1.00	0.00	4.00	191.53
oracle.indep	0.010	-	-	-	-	-
oracle.exch	0.003	-	-	-	-	-
oracle.ar1	0.003	-	-	-	-	-
PGEE.indep	0.011	0.00	0.83	0.17	4.00	2.15
PGEE.exch	0.004	0.00	0.33	0.67	4.00	4.23
PGEE.ar1	0.005	0.00	0.35	0.65	4.00	4.02

Table 2

Replicated study: correlated continuous reponse ($n = 200, p_n = 200$) with three different working correlation structures (independence, exchangeable, and AR(1)) and two different strengths of within cluster correlation

	MSE	U	O	EXACT	TP	FP
$\boldsymbol{\rho = 0.5}$						
GEE.indep	0.192	0.00	1.00	0.00	4.00	192.73
GEE.exch	253.64	0.00	1.00	0.00	4.00	195.83
GEE.ar1	0.191	0.00	1.00	0.00	4.00	192.78
oracle.indep	0.192	-	-	-	-	-
oracle.exch	0.191	-	-	-	-	-
oracle.ar1	0.191	-	-	-	-	-
PGEE.indep	0.189	0.00	0.76	0.24	4.00	1.56
PGEE.exch	4.545	0.23	0.00	0.77	3.59	0.00
PGEE.ar1	20.96	0.12	0.11	0.77	3.81	0.18
$\boldsymbol{\rho = 0.8}$						
GEE.indep	0.192	0.00	1.00	0.00	4.00	192.82
GEE.exch	104.25	0.00	1.00	0.00	4.00	195.85
GEE.ar1	0.191	0.00	1.00	0.00	4.00	191.71
oracle.indep	0.192	-	-	-	-	-
oracle.exch	0.831	-	-	-	-	-
oracle.ar1	0.191	-	-	-	-	-
PGEE.indep	0.189	0.00	0.41	0.59	4.00	0.49
PGEE.exch	2.586	0.70	0.00	0.30	2.75	0.00
PGEE.ar1	4.208	0.10	0.31	0.59	3.87	4.50