Penalized generalized estimating equations for high-dimensional longitudinal data analysis

D. Erik Boonstra and Logan Harris

Department of Biostatistics
The University of Iowa

May 8, 2023

Authors: Lan Wang, Jianhui Zhou, and Annie Qu

Outline

1. Introduction

2. Estimation and variable selection
   - Generalized estimating equations
   - Penalized GEEs

3. Simulations

4. The PGEE Package

5. Conclusions

Introduction

- High-dimensional longitudinal data consist of repeated measurements on a large number of covariates.
- The Framingham Hearty Study is an example of a large-scale study, where many covariates such as age, smoking status, cholesterol level, and blood pressure have been collected over a 75-year period.
- Another example is a yeast cell-cycle gene expression study, which includes 297 cell-cycle regulated genes and the binding probabilities for 96 transcription factors.

Estimation for longitudinal data

- Marginal models are an approach widely used for longitudinal data anlaysis, which are used when inference about the population effect is of interest instead of individual-level effect.
- These models provide estimates of regression parameters for linear or generalized linear models that characterize the relationship between the marginal expectation of the response and a set of explanatory variables.
- Methods have been developed for marginal models that separate the modeling of the regression of the response on the covariates from the modeling of the association among correlated observations on the response.

Variable selection

In the high-dimensional setting or when the number of variables is not large but various interaction effects are included, the number of covariates we must estimate in the marginal models becomes large.

- It often occurs that only a subset of these variables are revelant for modeling the response.
- Inclusion of extraneous variables can decrease the accuracy and efficiency of estimation and inference.
- Thus, variable selection is necesary.

Variable selection methods

With marginal models needing to include the intra-cluster correlation of the repeated observations, variable selection is challenging. Several methods have been proposed but all assume that the dimension of the predictors are fixed.

- Best subsets: QIC, generalized Mallow's C_p, BIC criterion based on the quadratic inference function (QIF)
- Continuous response: regularized non/semi-parametric modeling, variable selection for mixed-effects model proposed by Ni, Zhang, and Zhang (2009)
- Model-based variable selection: SCAD-penalized QIF, marginal generalized additive models, GEE based shrinkage estimator with artifical objective function

Notation

For marginal models, suppose the marginal expectation, $\mathbb{E}(\mathbf{Y}_i) = \boldsymbol{\mu_i}$ is a function of the explanatory variables, $h(\mu_{ij}) = \mathbf{x}_{ij}^\top \boldsymbol{\beta}$, where

- $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^\top$ for $i = 1, \ldots, N$ independent subjects,
- $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{in_i})^\top$ and $\mathbb{E}(Y_{ij}) = \mu_{ij}$,
- $h(\cdot)$ is a link function,
- $\mathbf{x}_{ij}$ is a vector of p covariates at observation j, and
- $var(Y_{ij}) = \phi v(\mu_{ij})$, where is $v(\cdot)$ is a known variance function and ϕ is the scale parameter to be estimated.

Effects of unknown covariance structure

Considering the true intra-cluster covariance structure of $\mathbf{Y}_i$ is often unknown, the maximum likelihood estimating equations cannot be solved.

$$\sum_{i=1}^{N} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \right)^{\top} var(\mathbf{Y}_i)^{-1}(\mathbf{Y}_i - \boldsymbol{\mu}_i) = 0 \tag{1}$$

Estimation for marginal models only

Liang and Zeger (1986) extended generalized linear models (GLMs) to correlated data setting and developed a marginal modeling estimation approach called *generalized estimating equations* (GEEs), which is based on the quasi-likelihood framework.

- For the response, the mean and variance structures must be specified to solve the estimating equations, along with the correlation structure for each cluster.
- Regression parameter estimators are consistent and asymptotically unbiased, assuming the mean structure is correctly specified.
- In large sample settings, the estimators are approximately multivariate normal, given that the number of clusters is large and the size of each cluster is relatively small.

Generalized estimating equations

With pre-specifying the correlation structure of each cluster, the estimating equations in (1) are modified to

$$\mathbf{S}(\beta, \boldsymbol{\alpha}) = \sum_{i=1}^{N} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \right)^{\top} \mathbf{A}_i(\phi)^{-1/2} \mathbf{R}_i(\boldsymbol{\alpha})^{-1} \mathbf{A}_i(\phi)^{-1/2} (\mathbf{Y}_i - \boldsymbol{\mu}_i), \qquad (2)$$

where $\mathbf{A}_i(\phi) = diag\,[var(Y_{i1}), \ldots, var(Y_{in_i})]$, $\mathbf{R}_i(\boldsymbol{\alpha})$ is the *working* correlation matrix, and $\boldsymbol{\alpha}$ is a vector of nuisance parameters.

Estimation and variable selection for marginal models

For simultaneous estimation and variable selection, *penalized generalized estimating equations* (PGEEs) may be used with estimating functions defined as

$$\mathbf{U}(\boldsymbol{\beta}) = \mathbf{S}(\boldsymbol{\beta}, \boldsymbol{\alpha}) + \mathbf{q}_\lambda(|\boldsymbol{\beta}|) sign(\boldsymbol{\beta})^\top, \tag{3}$$

where

$$\mathbf{q}_\lambda(|\boldsymbol{\beta}|) = [q_\lambda(|\beta_1|), \ldots, q_\lambda(|\beta_p|)]^\top \text{ and}$$
$$sign(\boldsymbol{\beta}) = [sign(\beta_1), \ldots, sign(\beta_p)]^\top$$

with $sign(t) = \mathbf{1}(t > 0) - \mathbf{1}(t < 0)$.

Remarks on PGEEs

- Since $\mathbf{U}(\boldsymbol{\beta})$ has discontinous points, an exact solution to $\mathbf{U}(\boldsymbol{\beta}) = 0$ may not exist. So, $\hat{\boldsymbol{\beta}}$ is the approximate solution to

$$\mathbf{U}(\hat{\boldsymbol{\beta}}) = o(a_n)$$

  for a sequence $a_n \to 0$.
- $q_\lambda(|\beta_j|)$ is zero for a large value of $|\beta_j|$ and is relatively large for a small value of $|\beta_j|$.
- Thus, $\mathbf{S}_j(\boldsymbol{\beta}, \boldsymbol{\alpha})$, the jth component of $\mathbf{S}(\boldsymbol{\beta}, \boldsymbol{\alpha})$, is not penalized if $|\beta_j|$ is large; while the penalty is large if $|\beta_j|$ is close (but not equal) to zero and therefore forces its estimator to be shrunken to zero.

PGEE estimation algorithm

The algorithm for solving penalized GEEs is an iterative algorithm that combines the minorization-maximization (MM) algorithm for non-convex penalties with the Newton-Raphson (NR) algorithm for the GEEs.

1. Determine a reasonable grid of values for λ.
2. Given a value of λ,
    - assign an initial value for $\boldsymbol{\beta}$,
    - compute $\mathbf{U}(\tilde{\boldsymbol{\beta}}), \mathbf{H}(\tilde{\boldsymbol{\beta}})$, and $\mathbf{E}(\tilde{\boldsymbol{\beta}})$ for current value of $\tilde{\boldsymbol{\beta}}$, which are expressed based on the MM and NR algorithms,
    - update current estimate of $\boldsymbol{\beta}$,
    - stop iteration if the convergence criterion is satisfied, and
    - compute the cross-validation (CV) value of λ.
3. Repeat step (2) for each λ and find the value λ_{min} that results in the smallest CV prediction error.
4. Select the $\hat{\boldsymbol{\beta}}$ that corresponds to λ_{min} and compute the covariance matrix of $\hat{\boldsymbol{\beta}}$.

See equations (5) for the mathemtical form of $\mathbf{H}(\tilde{\boldsymbol{\beta}})$ and $\mathbf{E}(\tilde{\boldsymbol{\beta}})$.

Simulation Set up

The authors consider a number of simulations to evaluate the performance of PGEEs.

- PGEEs are compared to two other model frameworks,
  - unpenalized GEEs and
  - oracle GEEs.
- Each of the models fit is compared across three different working correlation structures,
  - independence,
  - exchangeable, and
  - AR(1).
- Each simulation uses 100 generated datasets.
- λ is selected using 4-fold cross validation.
- Any coefficient below 10^{-3} is considered to be zero.

Evalutation

To evaluate the performance of each of the models, metrics are considered for both estimation accuracy and model selection performance.

- For estimation accuracy, MSE is estimated.
- For model selection performance, the proportion of
  - times under-selecting (U),
  - times over-selecting (O), and
  - times exactly selecting (EXACT)

  the covariates with nonzero coefficients is reported.
- Additionally, for model selection, the authors consider
  - the average number of selected covariates that correspond to the nonzero coefficients (TP), and
  - the average number of selected covariates that correspond to the zero coefficients (FP).

Correlated Normal Response

One specific scenario the authors consider is that of a correlated normal response. To simulate data, the authors generate responses from

$$Y_{ij} = \boldsymbol{X}_{ij}^T \boldsymbol{\beta} + \boldsymbol{\epsilon}_{ij}$$

where, n = 200, t = 4, and p = 200.
Additionally, $\boldsymbol{\beta}$ includes 4 non-zero coefficients (2.0, 3.0, 1.5, 2.0) and the respective covariates are generated with

- the first covariate being from a Bernoulli(0.5), and
- all other covariates from a MVN($\mathbf{0}$, $\boldsymbol{\Sigma}$), where $\boldsymbol{\Sigma}$ is AR(1) with $\sigma^2 = 1$ and $\rho = 0.5$.

Lastly, $\boldsymbol{\epsilon}_i$ are generated from a MVN($\mathbf{0}$, $\boldsymbol{\Sigma}$), where $\boldsymbol{\Sigma}$ is exchangeable with $\sigma^2 = 1$ and $\rho = \{0.5, 0.8\}$.

Correlated Normal Response

		$\rho = .5$					
		MSE	U	O	EXACT	TP	FP
Indep	GEE	0.568	0.00	1.00	0.00	4.00	193.02
	Oracle	0.009	-	-	-	-	-
	PGEE	0.009	0.00	0.85	0.15	4.00	2.02
Exch	GEE	0.381	0.00	1.00	0.00	4.00	192.45
	Oracle	0.006	-	-	-	-	-
	PGEE	0.008	0.00	0.33	0.67	4.00	3.30
AR(1)	GEE	0.458	0.00	1.00	0.00	4.00	192.66
	Oracle	0.007	-	-	-	-	-
	PGEE	0.008	0.00	0.38	0.62	4.00	3.00

Correlated Normal Response

		$\rho = .8$					
		MSE	U	O	EXACT	TP	FP
Indep	GEE	0.568	0.00	1.00	0.00	4.00	193.01
	Oracle	0.010	-	-	-	-	-
	PGEE	0.011	0.00	0.83	0.17	4.00	2.15
Exch	GEE	0.165	0.00	1.00	0.00	4.00	190.44
	Oracle	0.003	-	-	-	-	-
	PGEE	0.004	0.00	0.33	0.67	4.00	4.23
AR(1)	GEE	0.211	0.00	1.00	0.00	4.00	191.53
	Oracle	0.003	-	-	-	-	-
	PGEE	0.005	0.00	0.35	0.65	4.00	4.02

Take Aways

- PGEEs performs similarly to an oracle in terms of MSE, and provide a significant improvement over GEEs in this high dimensional setting.
- PGEEs are generally robust to misspecification of the working correlation structure.
  - They perform best under the true correlation structure.
  - They performs worst when independence is the working correlation structure.
- Differences among correlation structures are highlighted when correlation is higher. As intra-cluster correlation increases, models with correlation structure that can account for such correlation tend to improve whereas those which assume independence tend to suffer.

Overview

The **PGEE** package is straightforward to use and has three core functions,

- `CVfit` which determines the ideal value of λ using cross validation,
- `PGEE` which fits a PGEE with λ selected by `CVfit`, and
- `MGEE` which fits an unpenalized GEE.

When using the `PGEE` package, there are a couple details that the user should keep in mind. Specifically,

- the user must ensure that the intercept is not penalized, and
- the data must be sorted by `id`.

Additionally, the user should be aware that

- `CVfit` selects λ under a working independence correlation structure,
- `PGEE` only takes numeric variables, and
- `PGEE` applies the SCAD penalty.

Data

We will demonstrate the usage following the authors' example on the yeast cell-cycle data.

- The data contain 297 cell-cycle-regularized genes (Y) and the standardized binding probabilities of 96 transcription factors and time (X).
- Measurements were taken over two cell-cycle periods for a total of 18 time points.
- An individual analysis is done per stage of the cell cycle, and here the focus is on G1 (4 time points).
- The outcome (Y) is assumed to be normally distributed.

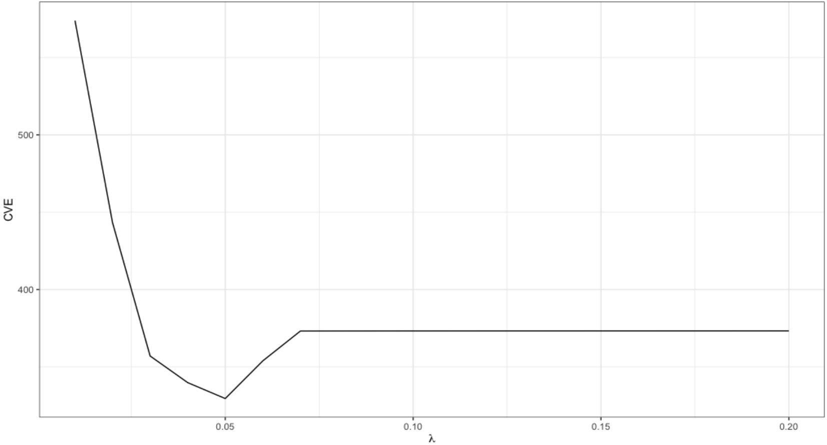
Goal: Identify TFs associated with gene expression levels

Run Cross Validation

```
library(PGEE); data(yeastG1) ## Load library and data
lambdas <- seq(0.2, 0.01, by = -0.01) ## Lambda seq

## Select optimal lambda
set.seed(050823)
cv <- CVfit(
  y ˜ . -id, id = id, data = yeastG1,
  family = gaussian(link = "identity"),
  scale.fix = TRUE, scale.value = 1, ## Dispersion
  fold = 10, lambda.vec = lambdas,
  pindex = c(1, 2) ## Don’t penalize int or time
)
```

Cross Validation Results



Fit PGEE

```
## Fit PGEE
fit <- PGEE(
  y ~ . -id, id = id, data = yeastG1,
  corstr = "AR-1",
  beta_int = NULL, ## Initial values set using a GLM
  lambda = cv$lam.opt, pindex = c(1, 2)
)
```

Obtain Results

```
## Get Non-zero Variables
coefs <- coef(summary(fit))
idx <- which(abs(coefs[, "Estimate"]) > 10^-3)
coefs[idx,] ## names(abs(ceofs[idx, "Estimate"]))
```

- This procedure selected 47 transcription factors, the following table provides an example of the output for `time` and two of the selected transcription factors.

Example Results

	Estimate	Naive SE	Naive Z	Robust SE	Robust Z
time	0.010	0.007	1.416	0.003	2.933
ABF1	-0.030	0.027	-1.110	0.012	-2.508
ACE2	0.008	0.020	0.390	0.007	1.165

Conclusions

- Penalized generalized estimating equations allow for simultaneous estimation and variable selection, which results in consistent and efficient model selection process.
- While this estimation procedure is robust to misspecification of the working correlation matrix, it may have some loss in efficiency when the correlation structure is misspecified. Additionally, PGEEs do not allow for one to determine the proper working correlation structure, as metric like QIC does.
- PGEEs provide significant improvement in estimation abilities over that of GEEs in high dimensional settings.
- Although PGEEs are relatively robust to specification of working correlation structure, there is a clear benefit to minimizing misspecification which becomes more apparent with increasing levels of correlation.

Questions

Thank you!

Appendix

PGEE estimation algorithm cont.

At the ℓth iteration of step (2) of the penalized GEEs algorithm,

$$\hat{\boldsymbol{\beta}}^{\ell} = \hat{\boldsymbol{\beta}}^{\ell-1} + \left[\mathbf{H}\left(\hat{\boldsymbol{\beta}}^{\ell-1}\right) + N\mathbf{E}\left(\hat{\boldsymbol{\beta}}^{\ell-1}\right)\right]^{-1}\left[\mathbf{S}\left(\hat{\boldsymbol{\beta}}^{\ell-1}\right) - N\mathbf{E}\left(\hat{\boldsymbol{\beta}}^{\ell-1}\right)\hat{\boldsymbol{\beta}}^{\ell-1}\right], \quad (4)$$

where

$$\mathbf{H}\left(\hat{\boldsymbol{\beta}}^{\ell-1}\right) = \sum_{i=1}^{N} \mathbf{X}_i^{\top}\mathbf{A}_i(\phi)^{-1/2}\mathbf{R}_i(\boldsymbol{\alpha})^{-1}\mathbf{A}_i(\phi)^{-1/2}\mathbf{X}_i \quad (5)$$

$$\mathbf{E}\left(\hat{\boldsymbol{\beta}}^{\ell-1}\right) = diag\left\{\frac{q_{\lambda}(|\hat{\beta}_1|+)}{\epsilon+|\hat{\beta}_1|}, \ldots, \frac{q_{\lambda}(|\hat{\beta}_p|+)}{\epsilon+|\hat{\beta}_p|}\right\} \quad (6)$$

$$(7)$$

with ϵ is a small value (e.g. 10^{-6}).

Cross-validation

The tuning parameter λ is selected based on K-fold cross-validation, where PGEE is fit under the working independence assumption using the training data and then evaluated the prediction error using the test data by $PE_{-k}(\lambda)$, which is defined as

$$PE_{-k}(\lambda) = \frac{1}{|N_{-k}|} \sum_{i \in N_{-k}} \frac{1}{n_i} \sum_{j=1}^{n_i} \left[Y_{ij} - h(\mathbf{x}_{ij}^\top \hat{\beta}) \right]^2$$

and the overall cross-validation error over the K subsamples is

$$CV(\lambda) = \frac{1}{K} \sum_{k=1}^{K} PE_{-k}(\lambda)$$

with $|N_{-k}|$ denoting the cardinality of N_{-k}.

Asymptotic results

Under regularity conditions, as the number of covariates p increases as N increase, and p can reach the same order as N, we have the following.

- If the working correlation structure is misspecified, the consistency of model select still holds (i.e., probability approaching one, PGEE correctly identifies the zero coefficients to be zero and the nonzero coefficients to be nonzero).
- The sandwich formula for covariance matrix is

$$cov(\hat{\beta}) \approx \left[\mathbf{H}\left(\hat{\beta}\right) + N\mathbf{E}\left(\hat{\beta}\right)\right]^{-1} \mathbf{M}(\hat{\beta}) \left[\mathbf{H}\left(\hat{\beta}\right) + N\mathbf{E}\left(\hat{\beta}\right)\right]^{-1},$$

where

$$\mathbf{M}(\hat{\beta}) = \sum_{i=1}^{N} \mathbf{X}_i^\top \mathbf{A}_i(\phi)^{-1/2} \mathbf{R}_i(\alpha)^{-1} \varepsilon_i \varepsilon_i^\top \mathbf{A}_i(\phi)^{-1/2} \mathbf{X}_i,$$

with $\varepsilon_i = \mathbf{A}_i(\phi)^{-1/2}(\mathbf{Y}_i - \boldsymbol{\mu}_i)$.

Scenario 1: Correlated Normal Response

Table 2 - $\rho = .5$

		β_1	β_2	β_3	β_4
Indep	Bias	0.002	0.000	0.002	0.003
	Est sd	0.067	0.041	0.046	0.041
	Em sd	0.065	0.038	0.043	0.041
	Cov Pr	96.00	96.00	97.00	97.00
Exch	Bias	0.000	0.002	0.001	0.000
	Est sd	0.051	0.032	0.036	0.032
	Em sd	0.053	0.030	0.036	0.031
	Cov Pr	95.00	96.00	93.00	97.00
AR(1)	Bias	0.001	0.001	0.001	0.000
	Est sd	0.054	0.034	0.038	0.034
	Em sd	0.056	0.032	0.038	0.034
	Cov Pr	95.00	97.00	95.00	95.00

Scenario 1: Correlated Normal Response

Table 2 - $\rho = .8$

		β_1	β_2	β_3	β_4
Indep	Bias	0.002	0.001	0.002	0.005
	Est sd	0.074	0.041	0.046	0.040
	Em sd	0.072	0.038	0.043	0.042
	Cov Pr	97.00	96.00	97.00	97.00
Exch	Bias	0.000	0.001	0.001	0.000
	Est sd	0.037	0.021	0.024	0.021
	Em sd	0.039	0.020	0.024	0.020
	Cov Pr	95.00	94.00	94.00	97.00
AR(1)	Bias	0.002	0.000	0.000	0.000
	Est sd	0.042	0.023	0.026	0.023
	Em sd	0.043	0.022	0.026	0.023
	Cov Pr	96.00	96.00	95.00	96.00

- The sandwich variance estimator performs well and provides reliable empirical coverage probabilities.

Scenario 2: Correlated Binary Response

To simulate data for this scenario, the authors generate responses from

$$logit(\pi_{ij}) = \boldsymbol{X}_{ij}^{T}\boldsymbol{\beta}$$

where n = 400, t = 10, and p = 50.
Additionally, $\boldsymbol{\beta}$ includes 3 non-zero coefficients (0.7, -0.7, -0.4), each with covariates generated from a Uniform(0, 1).
Finally, correlated binary responses were then generated using `mvtBinaryEP` with Exch($\rho = 0.4$) as the correlation structure.

Scenario 2: Correlated Binary Response

Table 3

		MSE	U	O	EXACT	TP	FP
Indep	GEE	0.635	0.00	1.00	0.00	3.00	46.51
	Oracle	0.027	-	-	-	-	-
	PGEE	0.111	0.28	0.32	0.40	2.72	0.93
Exch	GEE	0.421	0.00	1.00	0.00	3.00	46.60
	Oracle	0.018	-	-	-	-	-
	PGEE	0.049	0.03	0.37	0.60	2.97	1.05
AR(1)	GEE	0.576	0.00	1.00	0.00	3.00	46.59
	Oracle	0.025	-	-	-	-	-
	PGEE	0.081	0.09	0.37	0.54	2.91	1.33

Scenario 2: Correlated Binary Response

Table 4

		β_1	β_2	β_3
Indep	Bias	0.038	0.048	0.107
	Est sd	0.099	0.101	0.066
	Em sd	0.119	0.125	0.221
	Cov Pr	85.00	87.00	64.00
Exch	Bias	0.001	0.005	0.027
	Est sd	0.081	0.083	0.076
	Em sd	0.078	0.082	0.128
	Cov Pr	94.00	95.00	88.00
AR(1)	Bias	0.010	0.017	0.050
	Est sd	0.089	0.090	0.077
	Em sd	0.087	0.095	0.171
	Cov Pr	96.00	98.00	81.00

Scenario 2: Take Aways

The conclusions are similar to that of scenario 1, with a couple of notable differences.

- In general, this setting is more difficult.
- PGEE's MSE noticeably deviates from that of the oracle.
- PGEEs have a tendency to under select depending on the working correlation structure, likely due to β_3.
- β_3 gives PGEEs difficulty in general (Bias, SD).
- The differences under assumed correlation structures are more stark.