Penalized Generalized Estimating Equations for High-Dimensional Longitudinal Data Analysis

Authors: Lan Wang, Jianhui Zhou, Annie Qu

DOI: https://doi.org/10.1111/j.1541-0420.2011.01678.x

Presented by: Yao Chen

Introduction

- Penalized generalized estimating equations (GEE) for high-dimensional longitudinal data analysis
- Traditional GEE
- Penalty functions
- Theoretical properties
- Evaluation
    - Monte Carlo simulations
    - Real dataset application

Wang, L., Zhou, J. and Qu, A. (2012), Penalized Generalized Estimating Equations for High-Dimensional Longitudinal Data Analysis. Biometrics, 68: 353-360.

Longitudinal Data Analysis – Challenges

- High-dimensional longitudinal data
  - Repeated measurements on a large number of covariates over time
  - The number of variables (*p*) is much larger than the number of observations (*n*) , *p>>n*
  - Large-scale long-term health studies, gene expression experiments…
- Traditional Generalized Estimating Equations in high-dimensional settings
  - Variable selection
  - Parameter estimation

Wang, L., Zhou, J. and Qu, A. (2012), Penalized Generalized Estimating Equations for High-Dimensional Longitudinal Data Analysis. Biometrics, 68: 353-360.

Generalized Estimating Equations (GEE)

1. Correlated responses for each subject, *i*, measured at different time points: $\mathbf{Y_{ij}} = \left(\mathrm{Y}_{\mathrm{i1}}, \dots, \mathrm{Y}_{\mathrm{im_i}}\right)^{\mathbf{T}}$

   p_n-dimensional vector of covariates X_{ij} with $\mathbf{X_i} = \left(\mathbf{X_{i1}}, \dots, \mathbf{X_{im_i}}\right)^{\mathbf{T}}$

2. GEE estimates the population-average or marginal effect of the predictors on the outcome variable, rather than the subject-specific effect.
3. GEE can take into account the correlation of within-subject data (longitudinal studies) by specifying a working correlation matrix $\mathrm{R}(\tau)$ structure, e.g., independence, AR(1), exchangeable …
   - Misspecification can be problematic and affect efficiency of the parameter estimates.
   - To fix this, use GEE with the *Huber-White* ***"sandwich estimator"*** for robustness.
4. Likelihood-based methods are not available for usual statistical inference. GEE is a **quasi-likelihood method**. ONLY the first two moments, the **mean** and the **covariance** matter.
5. Unclear on how to perform model selection, as GEE is just an estimating procedure. There is no goodness-of-fit measure readily available

Wang, L., Zhou, J. and Qu, A. (2012), Penalized Generalized Estimating Equations for High-Dimensional Longitudinal Data Analysis. Biometrics, 68: 353-360.

GEE with Diverging Number of Covariates

Wang (2011) in paper "GEE Analysis of Clustered Binary Data with Diverging Number of Covariates," developed an asymptotic theory for GEE analysis of clustered binary data when the number of covariates grows to infinity with the number of clusters **(”large n, diverging p” framework**).

1. The existence, consistency, and asymptotic normality of the GEE estimator under appropriate regularity conditions
2. When the working correlation matrix is misspecified, ***the sandwich variance formula*** remains valid.

   An asymptotically valid confidence interval and Wald test for an estimable linear combination of the unknown parameters

   The accuracy of the asymptotic approximation is examined via numerical simulations.

Wang, L., Zhou, J. and Qu, A. (2012), Penalized Generalized Estimating Equations for High-Dimensional Longitudinal Data Analysis. Biometrics, 68: 353-360.

The Penalized GEE Method – Overview

- This new method builds upon the traditional Generalized Estimating Equations.
- Add p_n-dimensional vector of penalty functions $q_{\lambda_n}(|\beta_n|)$
  - Why not LASSO (L_1 penalty)? It does not satisfy the unbiasedness condition.
  - Choose the non-convex SCAD penalty $q_{\lambda_n}(\theta)$ as it achieves three desirable properties of variable selection : unbiasedness, sparsity and continuity
- Encourages sparsity, improve variable selection and reduce estimation bias.
- The turning parameter λ_n determines the amount of shrinkage

Wang, L., Zhou, J. and Qu, A. (2012), Penalized Generalized Estimating Equations for High-Dimensional Longitudinal Data Analysis. Biometrics, 68: 353-360.

The Penalized GEE Method – Algorithm

An iterative algorithm: combination of minorization-maximization algorithm for nonconvex penalty and the Newton-Raphson algorithm for the GEE.

1. Identifies the sparse set of covariates that are associated with the outcome variable.
2. Uses minorization-maximization algorithm to identify relevant covariates by approximating non-convex penalty with a convex function.
3. Uses Newton-Raphson algorithm to estimate the GEE parameters based on second-order derivatives of the likelihood function.

Useful for handling high-dimensional data and identifying relevant covariates.

Note that it requires careful selection of tuning parameters λ_n.

Wang, L., Zhou, J. and Qu, A. (2012), Penalized Generalized Estimating Equations for High-Dimensional Longitudinal Data Analysis. Biometrics, 68: 353-360.

Asymptotic Theory - Theorem

Oracle property of variable selection:

> the procedure estimates the true zero coefficient as zero with probability approaching one and estimates the nonzero coefficients as efficiently as if the true model is known in advance.

The number of covariates p_n increases as the number of clusters n increases, and p_n can reach the same order as n.

The consistency of model selection holds even if the working correlation structure is misspecified. Sandwich formula can also be obtained in PGEE to estimate the asymptotic covariance matrix.

Wang, L., Zhou, J. and Qu, A. (2012), Penalized Generalized Estimating Equations for High-Dimensional Longitudinal Data Analysis. Biometrics, 68: 353-360.

Monte Carlo Simulations – Design

- Penalized GEE (PGEE), unpenalized GEE and the oracle GEE
- Three different working correlation structures
    - Independence, exchangeable and AR(1)
- Set up
    - Generate 100 data sets for each set up in simulations
    - Run 30 iterations for each data set with the iterative algorithm to estimate β_{n0}
- Estimation accuracy
    - Mean square error (MSE)
    - The proportion pf times the methods under-selecting (U), over-selecting (O) and exactly selecting (EXACT)
    - The average false positives (FP) and the average true positives (TP)

Wang, L., Zhou, J. and Qu, A. (2012), Penalized Generalized Estimating Equations for High-Dimensional Longitudinal Data Analysis. Biometrics, 68: 353-360.

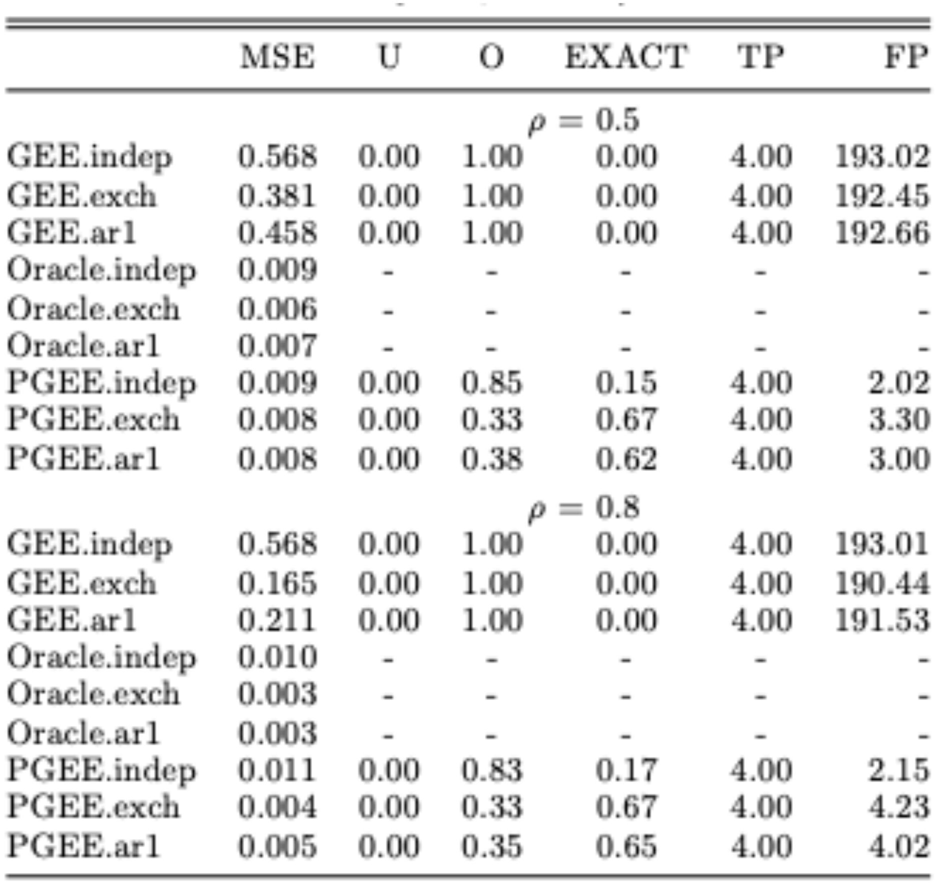
Monte Carlo Simulations – Results

Example 1 : correlated normal responses ($n = 200, p_n = 200$)

$$Y_{ij} = X_{ij}^T \beta + \epsilon_{ij}$$

The penalized GEE successfully reduces MSE, selects all covariates with nonzero coefficients and has a fairly small number of FPs.

$\rho = 0.5$	MSE	U	O	EXACT	TP	FP
GEE.indep	0.1916	0	1	0	4	192.73
GEE.exch	253.643	0	1	0	4	195.83
GEE.ar1	0.1912	0	1	0	4	192.78
Oracle.indep	0.1916	-	-	-	-	-
Oracle.exch	0.1909	-	-	-	-	-
Oracle.ar1	0.1912	-	-	-	-	-
PGEE.indep	0.1893	0	0.76	0.24	4	1.56
PGEE.exch	4.5451	0.23	0	0.77	3.59	0
PGEE.ar1	20.959	0.12	0.11	0.77	3.81	0.18



	MSE	U	O	EXACT	TP	FP
				$\rho = 0.5$		
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
				$\rho = 0.8$		
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

Wang, L., Zhou, J. and Qu, A. (2012), Penalized Generalized Estimating Equations for High-Dimensional Longitudinal Data Analysis. Biometrics, 68: 353-360.

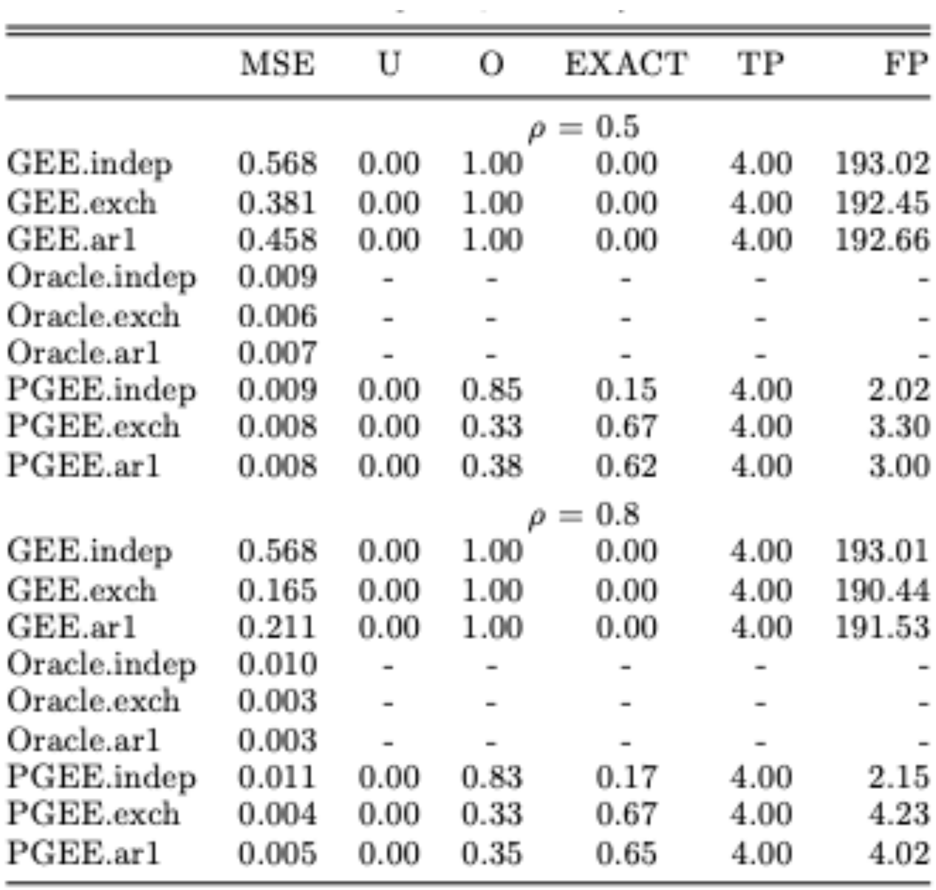
Monte Carlo Simulations – Results

Example 1 : correlated normal responses ($n = 200$, $p_n = 200$)

$$Y_{ij} = X_{ij}^T \beta + \epsilon_{ij}$$

The penalized GEE successfully reduces MSE, selects all covariates with nonzero coefficients and has a fairly small number of FPs.

$\rho = 0.8$	MSE	U	O	EXACT	TP	FP
GEE.indep	0.1921	0	1	0	4	192.82
GEE.exch	104.254	0	1	0	4	195.85
GEE.ar1	0.1912	0	1	0	4	191.71
Oracle.indep	0.1921	-	-	-	-	-
Oracle.exch	0.8306	-	-	-	-	-
Oracle.ar1	0.1914	-	-	-	-	-
PGEE.indep	0.1893	0	0.41	0.59	4	0.49
PGEE.exch	2.5857	0.7	0	0.3	2.75	0
PGEE.ar1	4.2079	0.1	0.31	0.59	3.87	4.5



	MSE	U	O	EXACT	TP	FP
			$\rho = 0.5$			
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
			$\rho = 0.8$			
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

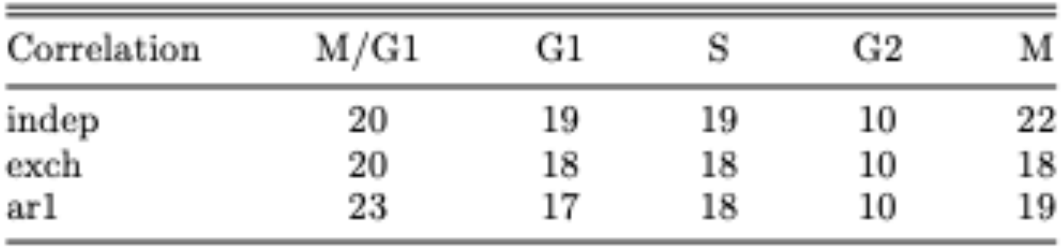
Example 2 for correlated binary responses is not replicated because the R package used was removed from CRAN repository.

Wang, L., Zhou, J. and Qu, A. (2012), Penalized Generalized Estimating Equations for High-Dimensional Longitudinal Data Analysis. Biometrics, 68: 353-360.

Real Data Application – Results

- Data: *yeast cell-cycle gene expression data* Genome-wide mRNA levels for 6178 yeast ORFs at 7-minute intervals for 119 minutes, which covers two cell-cycle periods for a total of 18 time points.
- Goal: Identify transcription factors (TFs) that influence gene expression levels at each stage of the cell process.
- At each of the five stages, the selected TFs were not sensitive to the choice of working correlation structure.
- Some of these selected TFs have already been confirmed by biological experiments using genome-wide binding method.
- Different TFs play important roles at different stages of the cell-cycle process, which has also been observed by the biologists.
    - The sets of TFs selected at different stages have only small overlaps.

Table: Number of TFs selected for each stage in the yeast cell-cycle process with the penalized GEE procedure.



Correlation	M/G1	G1	S	G2	M
indep	20	19	19	10	22
exch	20	18	18	10	18
ar1	23	17	18	10	19

Wang, L., Zhou, J. and Qu, A. (2012), Penalized Generalized Estimating Equations for High-Dimensional Longitudinal Data Analysis. Biometrics, 68: 353-360.

Advantages of Penalized GEE

1. Flexibility

   Penalized GEEs only require specifying the first two marginal moments and a working correlation structure

2. Model Selection Consistency

   The consistency of model selection holds even if the working correlation structure is misspecified.

3. Asymptotic Properties

   The number of covariates p_n increases as the number of clusters n increases and can reach the same order as n.

4. Performance in Practice

Wang, L., Zhou, J. and Qu, A. (2012), Penalized Generalized Estimating Equations for High-Dimensional Longitudinal Data Analysis. Biometrics, 68: 353-360.

Concluding Remarks – Potential Limitations

- Assumptions about working correlation structure
- Tendency to select an overfitted model in cross validation
- The tuning parameter selection may influence the practical performance
- Computational complexity with larger datasets (took me 18+ hours to only run R codes to replicate the simulations and real-world data analysis)

Wang, L., Zhou, J. and Qu, A. (2012), Penalized Generalized Estimating Equations for High-Dimensional Longitudinal Data Analysis. Biometrics, 68: 353-360.

Thank you and questions?

Wang, L., Zhou, J. and Qu, A. (2012), Penalized Generalized Estimating Equations for High-Dimensional Longitudinal Data Analysis. Biometrics, 68: 353-360.