Grouped variables selection via alternating direction method of multipliers in longitudinal analysis

Chengyu Zhang\textsuperscript{1}, Yihe Yang and Jianxin Pan

School of Mathematics, Sichuan University, China

\textsuperscript{1}Email: zhangchengyu94@163.com

Abstract. This work focuses on automatically and simultaneously selecting and estimating the coefficients of grouped variables in longitudinal data. It is motivated by a large cohort study, the UK North Staffordshire Osteoarthritis Project (NorStOP), which contains highly associated continuous, binary and ordinal variables. To select grouped variables associated with the target response efficiently, we propose a newly-developed generalized estimating equation (GEE) estimation. The innovation of this method is to study and implement GEE with group-lasso rather than traditional lasso which lacks consideration of group effects. We applied the proposed approach to the NorStOP data, and the results turned out to be reasonable.

1. Introduction

The aged people face many challenging problems to their quality of life, including a marked increase in the incidence of disease, higher levels of functional disability, etc. \cite{1} Obesity increases the incidence of disease in general, so it is necessary to care about the body mass index (BMI) of older people. With increasing age in people over 50 years, there is a noticeable increase in the rate of onset of pain that interferes with daily life. \cite{2} In the past decade, there were clinical studies investigating the health condition of older people, including the UK North Staffordshire Osteoarthritis Project (NorStOP) \cite{3} for the local area population with age over 50 years. The NorStOP study was lasting for six years. In addition to the subject baseline information such as age, gender, alcohol use, and income status, body mass index (BMI), pain interference (PI) degree, as well as anxiety and depression status were collected longitudinally. In particular, most of the variables including alcohol use, income status, PI degree, anxiety and depression status are ordinal, which may switch to the factor (dummy) variables in regression analysis. In longitudinal analysis, repeated measurements of continuous and discrete variables are routinely recorded from the same subject, \cite{4} and with the development of collecting and storing data, there are increasing requirements for new statistical approaches to deal with the complexity of longitudinal data. Existing methods in literature suggested to add a shrinkage penalty such as least absolute shrinkage and selection operator (LASSO) and smoothly clipped absolute deviation (SCAD) to the ordinary GEE, \cite{5-8} in order to select the important variables and estimate the corresponding coefficients simultaneously. These methods have fine performances in some fields, but may get stuck in longitudinal analysis, because they do not take grouped variables into account.

Group structure is common in longitudinal analysis. For example, switching an ordered variable to factor variable or approximating nonparametric function by basis function expansions attribute to group structure. The NorStOP data is a typical example. A natural objective in the clinical research is to select the important groups for a large number of explanatory variables in linear or generalized linear models, such as finding the important variables highly correlated with BMI in the NorStOP data.
A standard method for group selection in linear or generalized linear models is group-lasso [9] but it has shortcomings in many aspects. [10] Like most lasso based approaches, group-lasso leads to biased parameter estimates. More severely, the optimized coefficients by group-lasso do not have good properties, because they require that the variables in one group have to be orthogonal. [11] The main contribution of this work is to provide an ADMM [12] based algorithm to implement the grouped penalization, and to apply grouped penalization to longitudinal data. The proposed algorithm is simple and effective, and it does not require the orthogonality of the variables in one group. Additionally, the 1-norm penalty in group-lasso is replaced by the unbiased nonconvex penalty, [13] and group penalization is combined in generalized estimating equation (GEE) optimization. [14]

The rest of this paper is organized as follows. In section 2, we introduce the detailed algorithm of the GEE minimization with group penalization and conduct simulations. In section 3, we apply the proposed method to analyse the NorStOP data. Finally, we draw a brief conclusion of this work in section 4.

2. Model and method

2.1. Estimation procedure

Motivated by the NorStOP study, we propose a method to select the important grouped variables in the GEE framework. In this study, we assume that there are \( n \) subjects and \( n_i \) repeated measurements for the \( i \)th subject \((i = 1, 2, \ldots, n)\). Denote \( Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i}) \) with covariance matrix \( \text{cov}(Y_i) = \Sigma_i \). Besides, the grouped multivariate predictor \( X_i = (X_{iA_1}, X_{iA_2}, \ldots, X_{iA_J}) \) is associated with the mean of \( Y_i \) by the following equation:

\[
\mathbb{E}(Y_i | X_i) = g^{-1}(\sum_{j=1}^{J} \beta_{A_j}^T X_{iA_j}),
\]

where \( \beta = (\beta_{A_1}, \beta_{A_2}, \ldots, \beta_{A_J}) \) are the grouped regression coefficients, \( A_1, \ldots, A_J \) denote \( J \) groups of the covariates, and \( g() \) is a link function. Particularly, we assume that \( \beta \) can be divided into two parts \( \beta = (\beta_1, \beta_2) \) where \( \beta_2 = 0 \) and \( \beta_1 = (\beta_{A_1}, \beta_{A_2}, \ldots, \beta_{A_J}) \), indicating that only the covariate \( X_{i1} = (X_{IA_1}, \ldots, X_{IA_J}) \) are actually correlated with \( Y_i \). The main purpose of this work is to select significant variables from the unlabeled predictor \( X_i \) and to estimate the associated coefficients simultaneously and automatically.

The alternating direction method of multipliers (ADMM) is a Lagrangian based approach that is of attractive features for the minimization with (multivariate) penalization. Here we directly provide its application of group penalization. For details, one can refer to Boyd.et al (2011). [12] Consider the GEE minimization with group penalization:

\[
\min_{\beta} \frac{1}{2n} \sum_{i=1}^{n} (Y_i - \psi(\beta^T X_i))^T R_i (Y_i - \psi(\beta^T X_i)) + \sum_{j=1}^{J} P_{\lambda_j}(\|\beta_{A_j}\|_2)
\]

where \( \psi() = g^{-1}() \) and \( R_i \) is the inverse of the working covariance structure matrix in the original GEE minimization. And \( \lambda_j \) is adaptive by the scale of the groups as \( \lambda_j = \|X_{A_j}\|_2 / \lambda_0 \). Additionally, \( P_{\lambda}(\cdot) \) is a nonconvex penalty function such as smoothly clipped absolute deviation (SCAD) proposed by Fan and Li (2001). [13] To solve this problem, ADMM brings in a vector \( \mu \) of Lagrange multipliers associated with the constraint. The augmented Lagrangian object function is given by:

\[
Q(\beta, \theta, \mu) = \frac{1}{2} \sum_{i=1}^{n} (Y_i - \psi(\beta^T X_i))^T R_i (Y_i - \psi(\beta^T X_i))
\]
where \( \rho > 0 \) is a small fixed parameter. The ADMM method is to minimize the augmented Lagrangian object function \( Q(\beta, \theta, \mu) \) successively over \( \beta, \theta \) and \( \mu \), coming up with the updates:

\[
\beta^{t+1} = \arg\min_{\beta} Q(\beta, \theta^t, \mu^t),
\]

\[
\theta^{t+1} = \arg\min_{\theta} Q(\beta^{t+1}, \theta, \mu^t),
\]

\[
\mu^{t+1} = \mu^t + \rho(\beta^{t+1} - \theta^{t+1}),
\]

for iterations \( t = 0, 1, \ldots \). Under relatively mild conditions, this procedure converges to an optimal solution.

The above is the general updates, below we give the details. The update of \( \beta \) is implemented via Fisher-scoring algorithm. Specifically speaking,

\[
\frac{\partial Q}{\partial \beta} = -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial \psi}{\partial \beta} R_i \left( Y_i - \psi(X_i \beta) \right) + \mu + \rho(\beta - \theta),
\]

\[
\mathbb{E} \left( \frac{\partial^2 Q}{\partial \beta^2} \right) = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \psi}{\partial \beta} R_i \frac{\partial \psi}{\partial \beta} + \rho I,
\]

where \( I \) is an identity matrix. We minimize \( Q(\beta, \theta^t, \mu^t) \) via the ADMM based Fisher-scoring algorithm:

\[
\beta^{t,k+1} = \beta^{t,k} - s^k \mathbb{E} \left( \frac{\partial^2 Q}{\partial \beta^2} \right)^{-1} \frac{\partial Q}{\partial \beta} \beta^{t,k} \theta^t \mu^t
\]

such that \( \beta^{t,k} \to \beta^{t+1} \) as \( k \to +\infty \). Note that \( s^k \) is independent identically distributed as \( s^k \sim \text{Unif}(0,1) \), is the random step-size (which makes the update converge faster), and \( \frac{\partial Q}{\partial \beta} \beta^{t,k} \theta^t \mu^t \) denotes that the first and second order partial derivatives of \( \beta \) is evaluated at \( \beta^{t,k}, \theta^t \) and \( \mu^t \).

After updating \( \beta \), now we turn to \( \theta \). Unfortunately, the derivative of nonconvex penalty \( P(\cdot) \) is too complex, and so we adopt the local linear approximation (LLA) \([15]\) of nonconvex penalty:

\[
P_{\lambda_j} \left( \| \theta_{A_j} \|_2 \right) \approx P_{\lambda_j} \left( \| \beta_{A_j}^{t+1} \|_2 \right) + P'_{\lambda_j} \left( \| \beta_{A_j}^{t+1} \|_2 \right) \left( \| \theta_{A_j} \|_2 - \| \beta_{A_j}^{t+1} \|_2 \right) := \lambda_j^{t+1} \theta_{A_j} + c,
\]

where \( \lambda_j^{t+1} = P'_{\lambda_j} \left( \| \beta_{A_j}^{t+1} \|_2 \right) \) and \( c \) is the remained constant. Thus, the Karush-Kuhn-Tucker (KKT) conditions for \( \theta_{A_j} \) conditional on \( \beta_{A_j}^{t+1} \) and \( \mu_{A_j}^{t} \) can be given by

\[
\lambda_j^{t+1} z_{\theta_{A_j}} - \mu_{A_j}^{t} - \rho \left( \beta_{A_j}^{t+1} - \theta_{A_j} \right) = 0,
\]

where \( z_{\theta_{A_j}} = \frac{\theta_{A_j}}{\| \theta_{A_j} \|_2} \) if \( \| \theta_{A_j} \|_2 \neq 0 \), otherwise \( z_{\theta_{A_j}} \) is a vector satisfying \( \| z_{\theta_{A_j}} \|_2 \leq 1 \). After some simple algebra, we obtain

\[
\theta_{A_j}^{t+1} = \left( 1 - \frac{\lambda_j^{t+1}}{\| \beta_{A_j}^{t+1} + \mu_{A_j}^{t} \|_2} \right) \left( \beta_{A_j}^{t+1} + \frac{\mu_{A_j}^{t}}{\rho} \right).
\]
Note that equation (9) is explicit without assuming $X_{A_j}$ to be orthogonal in subgroup $A_j$, which greatly expands the application of group-penalization.

2.2. Simulation study
In this subsection, we conduct a simulation study to assess the performance of the proposed variables selection approach. We firstly generate the correlated response by copula model. Specifically, we set $F(\cdot; \mu_Y)$ to be the cumulative distribution function (c.d.f.) of continuous/discrete response $Y$, and $\Phi(\cdot)$ be the c.d.f. of standard normal distribution. Then, generate $Y$ by

- a. Generate $Z_i \sim \mathcal{N}(0, G_{n_i})$ where $\text{diag}(G_{n_i}) = (1, 1, \ldots, 1)^T$;
- b. Calculate $W_{ij} = \Phi(Z_{ij}), j = 1, 2, \ldots, n_i$;
- c. Calculate $Y_{ij} = F^{-1}(W_{ij}, \psi(X_{ij} \beta))$.

Additionally, we set $\beta = (\beta_0, \beta_{A_1}, \beta_{A_2}, \ldots, \beta_{A_m})$, where the true values are $\beta_0 = 1$, $\beta_{A_1} = (-1, -1)$, $\beta_{A_2} = (-1.5, 2)$, $\beta_{A_3} = (0.75, 1.5, 3)$, and other zeros. For the design matrix $X$, we generate $X_{ij} \sim \mathcal{U}(-1, 1)$ and fix them up when generating the response repeatedly.

In this simulation study, we fix $n_i \equiv 3$. Three common distributions, Gaussian, binary and poisson distribution are considered. We set $G_{n_i}$ to be a matrix of AR(1) structure with structural parameter $\rho = 0.5$, while the working correlation matrix $R_i$ is assumed to be the CS structure with structural parameter $\rho = 0.1$. We generate two data sets $n = 100$ and $n = 1000$, and run 300 times for each. The simulation results are presented Table 1, where GLM denotes the traditional method totally ignoring the group-in correlation, GEE can be regarded as our method with null penalty, and sparse GEE (SGEE) is our method.

| Table 1. Simulation for Gaussian, Binary and Poisson distributions via GLM, GEE and SGEE. |
|----------------------------------|-----------------|-----------------|
|                                  | Gaussian        | Binary          | Poisson         |
| n=100 GLM                        | 0.681           | 0.469           | 0.567           |
| MSE                              | 0.567           | 1.812           | 1.812           |
| Selection                        | -               | -               | -               |
| n=1000 GLM                       | 0.216           | 0.178           | 0.144           |
| MSE                              | 0.053           | 0.529           | 0.529           |
| Selection                        | -               | -               | -               |
| n=1000 GLM                       | 7.987           | 8.093           | 8.093           |
| MLE                              | -               | -               | -               |
| n=1000 GLM                       | 0.053           | 0.144           | 0.144           |
| MSE                              | 0.529           | 0.529           | 0.529           |
| Selection                        | -               | -               | -               |
| n=1000 GLM                       | 8.000           | 8.000           | 8.000           |
| MLE                              | -               | -               | -               |

From Table 1, we can clearly see the benefit brought by group-penalization. The mean-square-error (MSE) of our estimates is the smallest one compared with the other two methods in the three different distributions. Besides, our method achieves consistent variable selection, especially in binary and Poisson distributions. Table 1 shows that when the sample size is 1000, SGEE exactly selects eight variables to be nonzero. By contrary, when using the other two methods without group-penalization, the coefficient estimators are less accurate, and they fail to indicate the important grouped variables associated with the response.

3. Application

3.1. NorStOP study
This work is motivated by the UK North Staffordshire Osteoarthritis Project (NorStOP), a large cohort study of middle-aged and elderly people sampled from population in North Staffordshire in the UK between 2002 and 2008. The details of the NorStOP survey were shown in Thomas et al. (2007). [2]

All the sampled patients aged 50 and over were asked to fill in a Health Questionnaire in 2002 at baseline. Those patients who gave permission to be re-contacted were mailed similar questionnaires 3 and 6 years later. In the study, the response consists of body mass index (BMI) calculated from self-reported height and weight. The covariates include age, gender, alcohol use, adequacy of income,
perception of health and anxiety and depression ascertained at the Hospital Anxiety and Depression Scale (HADS). [16]

The aim of this research is to select significant variables which are notably correlated with BMI, so we consider the regression model of continuous response BMI, given by

\[
BMI_{ij} = \beta_0 + Age_{ij} \beta_1 + KAP_{ij} \beta_2 + ANX_{ij} \beta_{A3,1} + DEP_{ij} \beta_{A3,2} \\
+ Alcohol_{ij} \beta_{A4} + Income_{ij} \beta_{A5} + PI_{ij} \beta_{A6} + Locus_{ij} \beta_{A7} + \epsilon_{ij}
\]  

(13)

where KAP is the number of restrictions, Locus indicates the location, and ANX and DEP denote anxiety and depression index respectively. The scale of group 4-7 is \(\{\|A_4\|_0, \|A_5\|_0, \|A_6\|_0, \|A_7\|_0\} = \{4,3,4,4\}\) Particularly, we tie the variable ANX and DEP up because both of them are descriptions of emotion and are highly correlated. Without loss of generality, each variable (including the factor variables) is scaled to norm 1. The tuning parameter selection procedure is a sub-sampling validation with sub-sampling size \(s = 0.5n\) and sub-sampling times \(t = 500\). Besides, SCAD with clipped parameter \(a = 3\) is carried out to penalize the grouped coefficients. Moreover, the working structure matrix is set to be compound symmetry with the correlation coefficient \(c = 0.05\), augmented Lagrange parameters \(\rho = 0.7\), and the candidate set of tuning parameter is \(\{\lambda_1 = 0.5, \lambda_2 = 1, ..., \lambda_{20} = 10\}\).

![Figure 1. Scaled regression coefficient and their 95% confidence intervals.](image)

The results of the above regression are as follows. The tuning parameter selected via sub-sampling validation is \(\lambda_{10} = 5\) and the estimates of coefficients are \(\hat{\beta}_0 = 2566, \hat{\beta}_{A1} = -452, \hat{\beta}_{A2} = 16, \hat{\beta}_{A3} = (-95, 70), \hat{\beta}_{A4} = (0, 0, 0, 0), \hat{\beta}_{A5} = (-39, -72, -55), \hat{\beta}_{A6} = (31, 45, 58, 30), \hat{\beta}_{A7} = (0, 0, 0, 0).\) It is shown that the estimates of coefficients in the same group turn out be entirely zero or nonzero as we expected. Confidence intervals of the coefficients are also displayed in Figure 1. Note that, for visualization, each confidence interval of the corresponding coefficient is scaled to have norm.

**3.2. Discussion**

Alcohol and Locus are dropped out of the regression, indicating that BMI is independent of them. Especially, we are surprised that Alcohol has no influence on BMI because it seems that drinking makes people fat. We have tried different default factor of Alcohol but the results turned out to be the same, and this will be a topic worthy of further discussion.

Income has a negative influence on BMI, meaning that people who get high income are less likely to gain weight. This result supports our common impression about the relationship between income and
obesity. The middle and moneyed prefer high-quality food, exercise regularly and pay more attention to
health, so they are more likely to be free from obesity.

PI has a positive influence on BMI, owing to the dyskinesia on account of joint pain and stiffness. Besides, a strange result is that age has a negative influence on BMI. Anxiety and Depression have
opposite impacts on BMI. Anxious people are usually lack of appetite while most of the depressed ones
are crapulent. However, both of the two emotions are harmful to health.

4. Conclusions
This paper concentrates on the selection of grouped variables and estimation of their coefficients in
longitudinal data by GEE minimization. The main contribution is to propose an ADMM based algorithm
to solve the GEE minimization with group-penalization efficiently. Results from the simulation
illustrate the benefits of the proposed method. We apply the proposed method to the NorStop data and
efficiently select significant grouped variables that are associated with BMI. It helps to assess important
factors related to the health condition of the older people in UK, and then find solutions to improve their
quality of life. We will provide more technical details on algorithm convergence and asymptotic results
and more simulation results when presenting on the conference. There are a lot of work that can be done
further, for example we can investigate the performance of our approach for discrete responses.

References
[1] Gibson S J and Helme R D 2001 Age-related differences in pain perception and report Clin
Geriatr Med 17 433-456
[2] Thomas E, Mottram S and Peat G 2007 The effect of age on the onset of pain interference in a
general population of older adults Prospective findings from the North Staffordshire Osteoarthritis Project (NorSTOP) Pain 129 21-27
[3] Thomas E, Peat G, Harris L et al. 2004 The prevalence of pain and pain interference in a general
population of older adults cross-sectional findings from the North Staffordshire Osteoarthritis Project (NorSTOP) Pain 110 361-368
[4] Li Q, Pan J and Belcher J 2016 Bayesian inference for joint modelling of longitudinal continuous
binary and ordinal events Statistical Methods in Medical Research 25(6) 2521-2540
[5] Li G, Lai P and Lian H 2015 Variable selection and estimation for partially linear single-index
models with longitudinal data Statistics and Computing 25(3) 579-593
[6] Cho H and Qu A 2013 Model selection for correlated data with diverging number of parameters
Statistica Sinica 23 901-927
[7] Lian H, Liang H and Wang L 2014 Generalized additive partial linear models for clustered data
with diverging number of covariates using GEE Statistica Sinica 24(1) 173-196
[8] Wang L, Zhou J and Qu A 2012 Penalized generalized estimating equations for high-dimensional
longitudinal data analysis. Biometrics 68(2) 353-360
[9] Tibshirani R 1996 Regression shrinkage and selection via the lasso Journal of the Royal
Statistical Society: Series B (Methodology) 58 267-288
[10] Yuan M and Lin Y 2006 Model selection and estimation in regression with grouped variables
Journal of the Royal Statistical Society: Series B (Methodology) 68 49-67
[11] Tibshirani R, Wainwright M and Hastie T 2015 Statistical learning with sparsity the lasso and
generalizations[M] Chapman and Hall/CRC.
[12] Boyd S, Parikh N, Chu E, Peleato B and Eckstein J 2011 Distributed optimization and statistical
learning via the alternating direction method of multipliers. Foundations and Trends® in
Machine Learning 3 1-122
[13] Fan J and Li R 2001 Variable selection via nonconcave penalized likelihood and its oracle
properties Journal of the American Statistical Association 96 1348-1360
[14] Liang K Y and Zeger S L 1986 Longitudinal data analysis using generalized linear models
Biometrika 73 13-22
[15] Zou H and Li R 2008 One-step sparse estimates in nonconcave penalized likelihood models The
Annals of Statistics 36 1059

[16] Zogmond A S and Snaith R P 1983 The hospital anxiety and depression scale Acta Psychiatrca Scandinavica 67(6) 361-370