Data mining for longitudinal data under multicollinearity and time dependence using penalized generalized estimating equations[☆]

A. Blommaert [a,*], N. Hens [a,b], Ph. Beutels [a,c]

[a] Centre for Health Economics Research and Modeling Infectious Diseases, Centre for the Evaluation of Vaccination, Vaccine and Infectious Disease Institute (WHO Collaborating Centre), University of Antwerp, Antwerp, Belgium

[b] Interuniversity Institute for Biostatistics and statistical Bioinformatics, Hasselt University & Catholic University of Leuven, Hasselt, Belgium

[c] School of Public Health and Community Medicine, The University of New South Wales, Sydney, Australia

ARTICLE INFO

ABSTRACT

Penalized generalized estimating equations with Elastic Net or L2-Smoothly Clipped Absolute Deviation penalization are proposed to simultaneously select the most important variables and estimate their effects for longitudinal Gaussian data when multicollinearity is present. The method is able to consistently select and estimate the main effects even when strong correlations are present. In addition, the potential pitfall of time-dependent covariates is clarified. Both asymptotic theory and simulation results reveal the effectiveness of penalization as a data mining tool for longitudinal data, especially when a large number of variables is present. The method is illustrated by mining for the main determinants of life expectancy in Europe.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

Longitudinal data appear frequently in biomedical applications. Researchers are often confronted with the problem of determining the impact of different covariates on a response. Correct inference can be obtained by building an appropriate longitudinal model. Molenberghs and Verbeke (2005) distinguish three types of model families: marginal models, conditional models and subject-specific models. After the choice of the model family, an optimal set of predictors has to be selected. This can be a tedious task due to a large number of potential covariates. Including irrelevant covariates leads to inefficient inference. Therefore covariate selection is an important part of longitudinal model building, which is the main focus of this paper.

Variable selection in both the mixed model as a subject-specific model and generalized estimating equations as a marginal model will be briefly reviewed before turning to penalization methods within the generalized estimating equations framework.

Within the mixed model framework, Wu (2009) advised using significance testing or information criteria such as the Akaike information criterion or the Bayesian information criterion for the selection of fixed effects. Information criteria

---

have been further adapted to select both random and fixed effects in Jiang and Liu (2004) and Vaida and Blanchard (2005). Liu et al. (1999) generalized the idea of cross-validation to mixed models. When the number of covariates becomes large, employing a stepwise search can reduce the computational burden for the selection techniques above. Recently, Jiang et al. (2008) have suggested fence methods to put up a barrier between correct and incorrect mixed models.

In this paper we choose generalized estimating equations (GEE, Liang and Zeger, 1986; Zeger and Liang, 1986) as our inference framework instead of generalized linear mixed models (GLMM). The GEE approach yields population averaged effects by only specifying the first two moments of the outcome distribution. Its robustness against variance structure misspecification makes the GEE method well suited for our purpose of mean structure selection. Additionally, the problem of time-dependent covariates can be more easily clarified in the GEE context. The results provided in this paper are generalizable to linear mixed models as well, however this is not addressed here.

Despite the focus of GEE on mean structure estimation, appropriate covariate selection techniques are not well developed in this context. The standard practice for GEE model building is stepwise selection based on Wald-type tests (see for instance Diggle et al., 2002). More recently some general variable selection techniques have been adapted to the GEE framework. Pan (2001a) generalized the AIC to the GEE context based on the working independence assumption. Cantoni et al. (2005) suggested selection based on adequacy of prediction as measured by an adapted version of Mallow's C_p. In addition to these direct methods, more computationally intensive methods have been explored. Cantoni et al. (2007) combined cross-validation with a Markov Chain Monte Carlo based search. Alternatively, Pan (2001b) proposed minimizing a bootstrap smoothed cross-validation estimate of the expected predictive bias.

However, all of these methods lack the ability to properly deal with a large number of covariates. Because of the discrete nature of these selection methods, the resulting estimator can become unstable (Breiman, 1996). Moreover, complete subset comparison becomes computationally unfeasible when too many covariates are present, encouraging the use of a stepwise search. The gain in computation time by stepwise procedures comes at the price of suboptimal prediction performance and even higher instability.

In this paper we revisit the use of penalization within the GEE context to both reduce the computational burden and tackle the problem of instability. Indeed in ordinary regression and classification problems, penalization methods are well suited and often used for the task of variable selection and regularization. The Least Adaptive Shrinkage and Selection Operator (LASSO, Tibshirani, 1996), for example, is a penalization method which achieves both subset selection and parameter shrinkage. The continuous nature of the shrinkage leads to stable selection. The LASSO transforms the dimensionality of the subset selection problem into the selection of a single continuous tuning parameter. A major disadvantage of the LASSO is the potentially large bias induced by its shrinkage effect. The Smoothly Clipped Absolute Deviation penalty (SCAD, Fan and Li, 2001) is an adaptation to the LASSO which avoids unnecessary bias by using a different rate of penalization depending on the size of the coefficients. Smaller coefficients are penalized in the same manner as with the LASSO, while larger coefficients experience approximately no influence of the penalty.

Penalized generalized estimating equations (PGEE) were conceived by Fu (2003) as a framework in which these penalty methods can be applied in the longitudinal context. His asymptotic results were concentrated on bridge penalization (Frank and Friedman, 1993; Fu, 1998). Dziak (2006) and Dziak and Li (2007) extended this approach by using the SCAD penalty function. Even though their simulation studies display good performance of SCAD penalization for binomial data, we argue in Section 3 that their asymptotic results are limited to the Gaussian setting. Recently, Wang et al. (2011) have properly underpinned the SCAD penalized GEE with asymptotic theory for a response coming from the exponential family. Moreover their asymptotic results are constructed in a high dimensional-framework, allowing for the number of covariates p to diverge together with the number of clusters n. Assuming only that this divergence is of the same order as the increase in the number of clusters ($p = O(n)$), whereas in other aforementioned works p is assumed fixed.

In spite of the achievements of these authors, we believe that in mining for important variables in longitudinal data, two issues are commonplace, often overlooked and could be addressed better: multicollinearity and time-dependent covariates.

In order to deal with the first issue, multicollinearity, we suggest combining a sparse penalty function, namely the LASSO or the SCAD with a ridge part. In ordinary regression the elastic net (EN, Zou and Hastie, 2005) has been proposed as the combination of the LASSO and ridge regression. Recently the SCAD penalty has also been combined with a ridge part by Zeng (2009), an approach to which we refer hence as the $SCAD_{L2}$ penalty. The inclusion of a ridge part, adds the grouping effect to the resulting estimator. This means that highly correlated variables tend to be selected or omitted as a group.

The second issue, time-dependent covariates is often overlooked in this type of longitudinal analysis. Time dependence in generalized estimating equations will cause bias in the regression coefficients, unless either the cross-sectionality assumption is satisfied or the working independence matrix is used (Pepe and Anderson, 1994; Pan et al., 2000; Diggle et al., 2002).

In this paper we study EN and $SCAD_{L2}$ penalization within the framework of penalized generalized estimating equations with time-dependent covariates. We show how these methods deal with selection under multicollinearity using both asymptotic theory and simulation studies. We limit ourselves to the Gaussian setting with a fixed number of covariates and present avenues for generalization to the broader exponential family.

The remainder of the paper is organized as follows. In Section 2 we discuss the PGEE estimator with EN or $SCAD_{L2}$ penalty functions. As in Dziak (2006) we establish theory by turning to the equivalent penalized generalized least squares problem, but in contrast to Dziak (2006) argue that this is only possible for the Gaussian Case. The $SCAD_{L2}$ penalty function is shown to be convex under a condition on the tuning parameters. The equivalent penalized least square problem together