Ridge regression for longitudinal data with application to biomarkers

Melissa N. Eliot  
*University of Massachusetts at Amherst*

*Et al.*
Ridge regression for longitudinal data with application to biomarkers

Melissa N. Eliot*, Andrea S. Foulkes*, Muredach P. Reillya, and Jane Fergusona

*Division of Biostatistics and Epidemiology, UMass, Amherst; aDivision of Cardiology, UPenn School of Medicine

INTRODUCTION

Technological advances facilitating the acquisition of large arrays of biomarker data have led to new opportunities to study disease progression based on individual-level characteristics. This creates an analytical challenge, however, due to the large number of potentially informative markers, the high degrees of correlation among them, and changes that occur over time. To address these issues, we propose a mixed ridge estimator which integrates ridge regression into the mixed model framework in order to account for both the correlation induced by repeatedly measuring the outcome on each individual over time, as well as the potential high degree of correlation among predictor variables. An extension of the EM algorithm is described to account for unknown variance/covariance parameters. A simulation study is conducted to illustrate model performance and a data example is provided.

HYPOTHESIS

We predict that the mixed ridge estimator will result in somewhat biased coefficients with smaller standard deviations than those of the mixed model without ridge component. This will result in an improvement of power over the mixed model when correlations among predictors are sufficiently high, while type I error rates are maintained at about 0.05 for both methods.

METHODS

Mixed model with ridge component

Linear mixed effects model given by
\[ \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \mathbf{e} \]
where \( \mathbf{y} \) is an \( n \times 1 \) vector of responses, \( \mathbf{X} \) is a known design matrix of size \( n \times p \), \( \mathbf{Z} \) is a known design matrix of size \( n \times q \), \( \boldsymbol{\beta} \) is a \( p \times 1 \) vector of fixed effects, \( \mathbf{b} \) is a \( q \times 1 \) vector of random effects, \( \mathbf{e} \) is an \( n \times 1 \) vector of random errors.

Add ridge component to linear mixed effects model and solve
\[ \hat{\mathbf{b}}_{\text{ridge}} = \mathbf{Z}'(\mathbf{Z}'\mathbf{Z} + \lambda \mathbf{I}_q)^{-1}\mathbf{Z}'\mathbf{y} \]

where \( \lambda > 0 \) is the ridge parameter.

Hypothesis

The mixed model with ridge component is compared to the mixed model without ridge component using a simulation study. The hypothesis is that the mixed model with ridge component will have smaller standard errors than the mixed model without ridge component when correlations among predictors are sufficiently high, while type I error rates are maintained at about 0.05 for both methods.

METHODS cont.

Consider the setting in which the variance parameters \( \sigma^2 \) and \( \tau^2 \) are unknown. Consider an EM algorithm to estimate the variance parameters. The algorithm proceeds as follows:

1. **Initialization:** Set \( \hat{\sigma}^2 = \text{var} \left( \mathbf{y} \right) / n \) and \( \hat{\tau}^2 = \text{var} \left( \mathbf{y} \right) / n \).

2. **E-step:** For each \( i \) from 1 to \( n \), set
\[ \hat{\mathbf{Z}}_i = \mathbf{Z}_i \mathbf{b}_i + \mathbf{e}_i \]

3. **M-step:** Set
\[ \hat{\sigma}^2 = \frac{n}{n-1} \mathbf{y}' \hat{\mathbf{Z}} \hat{\mathbf{Z}}' \mathbf{y} \]
\[ \hat{\tau}^2 = \frac{n}{n-1} \mathbf{y}' \hat{\mathbf{Z}} \hat{\mathbf{Z}}' \mathbf{y} - \hat{\sigma}^2 \]

4. Repeat steps 2 and 3.

Testing

To determine the significance of each predictor variable, we calculate Wald statistics by dividing each estimated coefficient by the square root of its variance. Since an EM algorithm is used, we use Louis' formula (Louis 1982) to determine variance. Finally, Westfall and Young's (1993) free step-down resampling approach is applied to adjust for multiple testing. All tests are two-tailed.

Simulation Study

A simulation study is performed to characterize the relative performances of mixed ridge regression and the usual mixed effects modeling approach in the context of multiple, correlated predictors. For simplicity we assume repeatedly measured outcomes and only baseline predictors. We let \( n = 500 \) measurements per subject and generate data according to the mixed-effects model, where \( \beta = (0.0, 0.1, 0.2, 0.3, 0.4, 0.5) \). Each predictor is assumed to arise from a normal distribution with mean \( 5 \) and variance \( 1 \). The correlation between predictor variables \( (\rho) \) takes on values between 0 and 0.99. Starting values for variance components are derived from the mixed model, after p-values are adjusted for multiple testing using the Westfall and Young approach. The data arising from this study is longitudinal with predictors with correlation coefficients of up to 0.95, which indicates that ridge regression is appropriate. We perform the analysis using the statistical software R and compare our results with the mixed model.

RESULTS

MR outperforms the mixed model without ridge component when correlations among predictor variables are sufficiently large. The simulation study shows that when correlations are greater than about 0.80, power of MR is higher than that of the mixed model without a significant increase in type I error rate. At lower correlations, MR works just as well as the mixed model. The GENE study data set included predictors with correlation coefficients as high as 0.95, and subjects were measured 2 to 4 times each. Due to the high correlation, mixed modeling resulted in inflated variances of coefficients, and thus low power. The MR approach identified APOE as significantly associated with BP over time while the usual mixed modeling approach was unable to detect this association.

CONCLUSIONS

Support for this research was provided by NIH award R01HL107196.