May 20th, 5:00 PM - 7:00 PM

Ridge regression for longitudinal data with application to biomarkers

Melissa N. Eliot
*University of Massachusetts at Amherst*

Andrea S. Foulkes
*University of Massachusetts at Amherst*

Muredach P. Reilly
*University of Pennsylvania School of Medicine*

*See next page for additional authors*

Follow this and additional works at: [https://escholarship.umassmed.edu/cts_retreat](https://escholarship.umassmed.edu/cts_retreat)

Part of the [Biostatistics Commons](https://escholarship.umassmed.edu/biostatistics), and the [Systems Biology Commons](https://escholarship.umassmed.edu/systemsbiology)

---


This material is brought to you by eScholarship@UMMS. It has been accepted for inclusion in UMass Center for Clinical and Translational Science Research Retreat by an authorized administrator of eScholarship@UMMS. For more information, please contact Lisa.Palmer@umassmed.edu.
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 opportunities for 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

Motivation

Problem: Predictor variables highly correlated → no unique solution to least squares and maximum likelihood estimates, or resulting coefficient estimates have inflated variances resulting in low predictive accuracy.

Solution: Ridge regression for longitudinal data, which we call the mixed ridge (MR) estimator.

Mixed ridge model

Linear mixed effects model given by

\[ Y = X' \beta + \epsilon \]

where \( Y \) is the response, \( X \) is the design matrix of predictors, \( \beta \) is the vector of fixed effects, and \( \epsilon \) is the error term.

Add ridge component to linear mixed effects model and solve

\[ \hat{\beta}_{MR} = \text{arg min} \{ (Y - X' \hat{\beta})' V^{-1} (Y - X' \hat{\beta}) + \rho \hat{\beta}' \hat{\beta} \} \]

for \( \hat{\beta}_{MR} \), where \( \rho \) is the ridge parameter.

Solution to (1) is given by

\[ \hat{\beta}_{MR} = (X' V^{-1} X + \rho I)^{-1} X' V^{-1} Y \]  

(2)

Additionally, \( \text{Var}(\hat{\beta}_{MR}) = (X' V^{-1} X + \rho I)^{-1} \) and

\[ \text{Bias}(\hat{\beta}_{MR}) = (X' V^{-1} X)^{-1} \text{Bias} \]

Using the GCV method proposed by Craven and Wahba (1979) we can estimate \( \rho \) by solving

\[ \hat{\rho} = \text{arg min} \{ (Y - X' \hat{\beta})' V^{-1} (Y - X' \hat{\beta}) + \rho \hat{\beta}' \hat{\beta} \} \]

for \( \hat{\rho} \), where \( \hat{\beta} \) is the fitted ridge regression coefficient.

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, and type I error rates are roughly the same. Table 1 compares MR and mixed model for data example. At the 0.05 level, MR finds 2 variables to be significant, compared with 1 variable for the mixed model. Figures 1 and 2 present QQ plots of t-statistics for MR and mixed model, respectively.

CONCLUSIONS

The data arising from this study is longitudinal with predictors with correlation coefficients of up to 0.95, which indicates ridge regression is appropriate. We perform the analysis using the lipid measurements at times 0, 6, 12, and 24 hours as predictors and systolic blood pressure as outcome. Because we expect that change in systolic blood pressure between times 0, 6, 12, and 12-24 hours is piecewise-linear, we use linear splines (Fitzmaurice, et. al) with “knots” or change points at times 6 and 12. Also included are random within-subject effects for intercept and slope. MR is compared with the mixed model, after p-values are adjusted for multiple testing using the Westfall and Young approach.

ACKNOWLEDGEMENTS

Support for this research was provided by NIH award R01HL107196.