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  
Part of the [Biostatistics Commons](https://escholarship.umassmed.edu/biostatistics_commons), and the [Systems Biology Commons](https://escholarship.umassmed.edu/systems_biology_commons)
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. However, due to the large number of potentially informative markers, the high degrees of correlation among them, and changes that occur over time, it is challenging 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 degrees 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 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 degrees 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.

Mixed ridge model

Linear mixed effects model given by
\[ Y = X^-1 X \hat{b} + X^-1 e, \]
where \( X \) denotes a matrix of indicator predictors, \( e \) are individual error terms, \( b \) are fixed effects, and \( Y \) is the outcome variable. The correlation between predictor variables is handled by the mixed effects model and solved by solving
\[ (X^-1 X + n \lambda) \hat{b} = X^-1 Y, \]
where \( \lambda \) is a tuning parameter. The solution to (1) is given by
\[ \hat{b} = (X^-1 X + n \lambda)^{-1} X^-1 Y. \]

Add ridge component to linear mixed effects model and solve
\[ \hat{b} = (X^-1 X - \lambda) \hat{b} + \lambda X^-1 Y, \]
where \( \lambda \) is a tuning parameter. The solution to (1) is given by
\[ \hat{b} = (X^-1 X + n \lambda)^{-1} X^-1 Y. \]

EM algorithm

Consider the setting in which the variance parameters \( \phi \) are unknown. We propose an extension of the expectation-maximization (EM) algorithm described by Laird and Ware (1982) that includes an additional step for estimation of the ridge component. The algorithm proceeds as follows:

1. (Expect step) Initialize \( \phi \) and solve (2) for \( \hat{b} \) and \( \hat{y} \) given by
\[ \hat{b} = (X^-1 X - \lambda) \hat{b} + \lambda X^-1 Y, \]
\[ \hat{y} = \hat{b} + X \hat{e}. \]

2. (Max step) Solve for \( \phi \) with \( \hat{b} \) and \( \hat{y} \) fixed.

3. Update \( \phi \) using equation (3) and set
\[ \phi = \sum_i \hat{e}_i^T \hat{e}_i \]

4. Repeat steps (1) to (3) a large number of times and select a convergence criterion.

Data Example

The GENE (Genetics of Evoked-Responses to Niacin and Endotoxemia) study is an ongoing trial designed to characterize the effects of genetic factors on the response to niacin therapy and endotoxemia. Healthy volunteers were assigned to an endotoxin, which produces a mild-inflammatory response that can last from 6-8 hours. At certain time points during a 24-hour period, vital signs such as blood pressure and temperature were measured, as well as TNF-alpha, TGF-beta, IL-6, IL-8, IL-10, and CRP.

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. MR is compared with the mixed model, after p-values are adjusted for multiple testing using the Westfall and Young approach.

CONCLUSIONS

MR outperforms the mixed model without ridge component when correlations among predictor variables are sufficiently large. The simulation study shows that the correlation between predictor variables is handled by the mixed effects model and solved by solving
\[ (X^-1 X + n \lambda)^{-1} X^-1 Y. \]

EM algorithm

Consider the setting in which the variance parameters \( \phi \) are unknown. We propose an extension of the expectation-maximization (EM) algorithm described by Laird and Ware (1982) that includes an additional step for estimation of the ridge component. The algorithm proceeds as follows:

1. (Expect step) Initialize \( \phi \) and solve (2) for \( \hat{b} \) and \( \hat{y} \) given by
\[ \hat{b} = (X^-1 X - \lambda) \hat{b} + \lambda X^-1 Y, \]
\[ \hat{y} = \hat{b} + X \hat{e}. \]

2. (Max step) Solve for \( \phi \) with \( \hat{b} \) and \( \hat{y} \) fixed.

3. Update \( \phi \) using equation (3) and set
\[ \phi = \sum_i \hat{e}_i^T \hat{e}_i \]

4. Repeat steps (1) to (3) a large number of times and select a convergence criterion.

Data Example

The GENE (Genetics of Evoked-Responses to Niacin and Endotoxemia) study is an ongoing trial designed to characterize the effects of genetic factors on the response to niacin therapy and endotoxemia. Healthy volunteers were assigned to an endotoxin, which produces a mild-inflammatory response that can last from 6-8 hours. At certain time points during a 24-hour period, vital signs such as blood pressure and temperature were measured, as well as TNF-alpha, TGF-beta, IL-6, IL-8, IL-10, and CRP.

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. MR is compared with the mixed model, after p-values are adjusted for multiple testing using the Westfall and Young approach.

CONCLUSIONS

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. MR is compared with the mixed model, after p-values are adjusted for multiple testing using the Westfall and Young approach.