Ridge regression for longitudinal data with application to biomarkers

Melissa N. Eliot*, Andrea S. Foulkes*, Muredach P. Reilly[a], and Jane Ferguson[a]

*Division of Biostatistics and Epidemiology, UMass, Amherst; [a]Division 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

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 precision.

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

Mixed ridge model

Linear mixed effects model given by $\mathbf{y}_i = \mathbf{X}_i\beta + \mathbf{z}_i^T b_i + \epsilon_i$, $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{in_i})^T$ for individual $i = 1, \ldots, n$ with n_i observations, $b_i \sim MVN(0, \mathbf{D})$, $\epsilon_i \sim MVN(0, \sigma_e^2 \mathbf{I}_{n_i \times n_i})$

Then $\widehat{\beta} = (\mathbf{X}^T V^{-1} \mathbf{X})^{-1} \mathbf{X}^T V^{-1} \mathbf{Y}$ where V is the variance of Y.

Add ridge component to linear mixed effects model and solve

$$\widehat{\beta}_{MR} = \arg\min_{\beta} \left\{ (\mathbf{Y} - \mathbf{X}\beta)^T V^{-1} (\mathbf{Y} - \mathbf{X}\beta) + \lambda \beta^T \beta \right\} \qquad (1)$$

Solution to (1) is given by

$$\widehat{\beta}_{MR} = (\mathbf{X}^T V^{-1} \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T V^{-1} \mathbf{Y} \qquad (2)$$

Additionally, $Var(\widehat{\beta}_{MR}) = (\mathbf{X}^T V^{-1} \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T V^{-1} \mathbf{X} \, (\mathbf{X}^T V^{-1} \mathbf{X} + \lambda \mathbf{I})^{-1}$ and

$\widehat{b} = E[b|\mathbf{Y}] = D\mathbf{Z}^T V^{-1}(\mathbf{Y} - \mathbf{X}\widehat{\beta}_{MR}) = D\mathbf{Z}^T V^{-1} \left(\mathbf{I} - \mathbf{X} \left[(\mathbf{X}' V^{-1} \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T V^{-1} \right] \right) \mathbf{Y}$

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

$$\widehat{\lambda} = \arg\min_{\lambda} \left\{ n^{-1} (\mathbf{Y} - \widehat{\mathbf{Y}})^T (\mathbf{Y} - \widehat{\mathbf{Y}}) / (1 - tr(\mathbf{S})/n)^2 \right\} \qquad (3)$$

METHODS cont.

Where $\mathbf{S} = \mathbf{X}(\mathbf{X}^T V^{-1} \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T V^{-1} + \mathbf{Z} D \mathbf{Z}^T (\mathbf{I} - \mathbf{X}[(\mathbf{X}^T V^{-1} \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T V^{-1}])$.

EM algorithm

Consider the setting in which the variance parameters $\theta = (\sigma, V)$ 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. (**E-step**) Initialize $\widehat{\theta}^{(t)} = \theta_0$ and $\widehat{\lambda}^{(t)} = \lambda_0$. Solve for $\widehat{\beta}_{\mathrm{MR}}^{(t)}$ and the sufficient statistics $\widehat{t}_1^{(t)}$ and $\widehat{t}_2^{(t)}$ given by:

$$\widehat{t}_1^{(t)} = E\left(\sum_{i=1}^{n} \epsilon_i^T \epsilon_i | \mathbf{Y}_i, \widehat{\beta}_{\mathrm{MR}}^{(t)}, \widehat{\theta}^{(t)} \right)$$

$$\widehat{t}_2^{(t)} = E\left(\sum_{i=1}^{n} b_i \, b_i^T | \mathbf{Y}_i, \widehat{\beta}_{\mathrm{MR}}^{(t)}, \widehat{\theta}^{(t)} \right)$$

2. (**M-Step**) Solve for $\widehat{\theta}^{(t+1)}$ where $\widehat{\sigma}^{2(t+1)} = \widehat{t}_1^{(t)}/N$, $\widehat{D}^{(t+1)} = \widehat{t}_2^{(t)}/n$, $N = \sum_{i=1}^{n} n_i$ and n is the number of individuals in our sample, and let $\widehat{V}^{(t+1)} = \mathbf{Z}\widehat{D}^{(t+1)}\mathbf{Z}^T + \widehat{\sigma}^{2(t+1)}\mathbf{I}$
3. Update $\widehat{\lambda}^{(t+1)}$ using equation (3) and let

$$\widehat{\beta}_{\mathrm{MR}}^{(t+1)} = \left(\mathbf{X}^T \widehat{V}^{-1(t+1)} \mathbf{X} + \widehat{\lambda}^{(t+1)} \mathbf{I} \right)^{-1} \mathbf{X}^T \widehat{V}^{-1(t+1)} \mathbf{Y}$$

4. Repeat Steps (1) – (3) a large number of times and until a convergence criterion is met.

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.

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 $n_i = 4$ measurements per subject and generate data according to the mixed-effects model, where $\beta = (0.0, 0.0, 0.2, 0.4, 0.6)$, $b_i \sim N(0, 0.6)$, and $\epsilon_{ijk} \sim N(0, 1)$. Each predictor is assumed to arise from a Normal distribution with mean 5 and variance 1. The correlation between predictor variables (ρ) takes on values between 0 and 0.99. Starting values for variance components are derived from fitting a mixed model with no ridge component. In total *M* = 500 simulations are conducted for each correlation with sample sizes of *n* = 500 individuals.

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 endotoxin. Healthy volunteers were given 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 Tumor necrosis factor-alpha (TNF), Apolipoprotein A1 (Apo-A1), Apolipoprotein B (Apo-B), Cholesterol (Chol) High-density lipoprotein (HDL), Low-density lipoprotein (cLDL), Phospholipids (Phos), and Triglycerides (Tri).

The data arising from this study is longitudinal with predictors with correlation coefficients of up to 0.95, which indicates that MR 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, 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.

RESULTS

Table 1

		Mixed		Mixed Ridge	
ρ	β	Estimate (sd)	Power	Estimate (sd)	Power
0.0	0.0	0.002 (0.031)		0.005 (0.023)	
	0.0	-0.001 (0.031)		0.002 (0.023)	
	0.2	0.196 (0.031)	1.0	0.197 (0.023)	1.0
	0.4	0.399 (0.031)	1.0	0.397 (0.023)	1.0
	0.6	0.602 (0.031)	1.0	0.597 (0.023)	1.0
0.2	0.0	-0.004 (0.031)		0.0026 (0.023)	
	0.0	-0.004 (0.031)		0.0020 (0.023)	
	0.2	0.197 (0.034)	1.0	0.203 (0.025)	1.0
	0.4	0.402 (0.034)	1.0	0.398 (0.025)	1.0
	0.6	0.601 (0.035)	1.0	0.595 (0.025)	1.0
0.4	0.0	-0.004 (0.030)		-0.002 (0.022)	
	0.0	0.004 (0.030)		0.006 (0.022)	
	0.2	0.199 (0.039)	1.0	0.200 (0.029)	1.0
	0.4	0.402 (0.039)	1.0	0.400 (0.029)	1.0
	0.6	0.602 (0.039)	1.0	0.595 (0.029)	1.0
0.6	0.0	0.004 (0.029)		0.005 (0.022)	
	0.0	-0.001 (0.030)		-0.001 (0.022)	
	0.2	0.191 (0.047)	0.98	0.206 (0.031)	1.0
	0.4	0.401 (0.047)	1.0	0.398 (0.031)	1.0
	0.6	0.603 (0.046)	1.0	0.592 (0.031)	1.0
0.8	0.0	0.003 (0.029)		0.005 (0.022)	
	0.0	-0.002 (0.029)		0.000 (0.022)	
	0.2	0.198 (0.065)	0.85	0.211 (0.047)	0.90
	0.4	0.394 (0.065)	1.0	0.393 (0.046)	1.0
	0.6	0.607 (0.064)	1.0	0.591 (0.046)	1.0
0.9	0.0	0.003 (0.029)		0.005 (0.021)	
	0.0	-0.002 (0.029)		0.000 (0.021)	
	0.2	0.204 (0.091)	0.52	0.227 (0.062)	0.74
	0.4	0.399 (0.091)	0.99	0.399 (0.062)	1.0
	0.6	0.596 (0.091)	1.0	0.568 (0.062)	1.0
0.99	0.0	0.003 (0.029)		0.005 (0.021)	
	0.0	-0.002 (0.029)		0.000 (0.021)	
	0.2	0.210 (0.288)	0.16	0.254 (0.143)	0.61
	0.4	0.402 (0.288)	0.28	0.404 (0.143)	0.73
	0.6	0.587 (0.286)	0.51	0.536 (0.142)	0.90

Table 2

	Mixed		Mixed Ridge	
Variable name	Estimate (sd)	t-statistic (p-value)	Estimate (sd)	t-statistic (p-value)
Time	0.46 (0.16)	2.90 (0.043)	0.47 (0.12)	3.99 (0.006)
APOb	-0.18 (0.09)	-1.94 (0.271)	-0.24 (0.07)	-3.59 (0.018)
Phos	-0.05 (0.03)	-1.35 (0.336)	-0.06 (0.02)	-2.67 (0.104)
t_1^* Spline	-0.40 (0.26)	-1.55 (0.387)	-0.40 (0.20)	-2.00 (0.366)
APOa	0.08 (0.05)	1.79 (0.271)	0.07 (0.04)	1.96 (0.366)
Chol	0.49 (1.20)	0.41 (0.769)	1.16 (0.94)	1.23 (0.696)
HDL	-0.55 (1.20)	-0.46 (0.769)	-1.16 (0.94)	-1.23 (0.696)
t_2^* Spline	-0.17 (0.18)	-0.93 (0.649)	-0.18 (0.15)	-1.15 (0.696)
Trig	-0.08 (0.24)	-0.34 (0.769)	-0.21 (0.19)	-1.09 (0.696)
cLDL	-0.35 (1.20)	-0.29 (0.769)	-0.98 (0.94)	-1.04 (0.696)
TNF	0.15 (0.20)	0.75 (0.712)	0.11 (0.13)	0.84 (0.696)

Figure 1

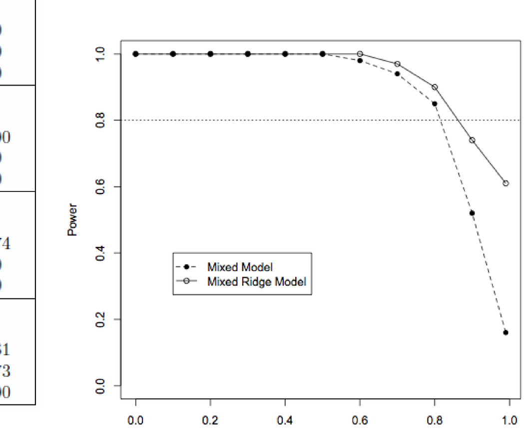

Figure 2

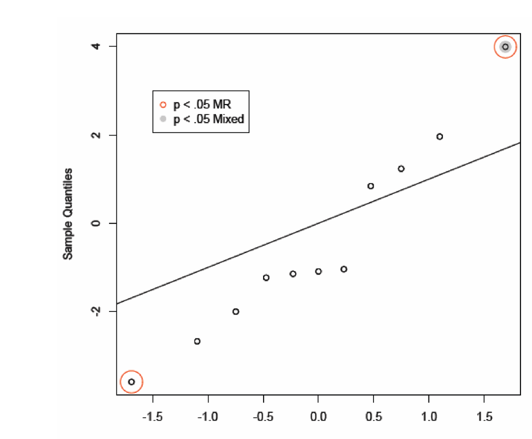

Table 1: Comparison of MR and Mixed model for simulation study. As correlation among columns increases, power of mixed model decreases more rapidly than that of MR. MR coefficients tend to have smaller variance and slight bias; however, type I error rates are roughly the same. **Figure 1**: Plot of power over correlations when β=0.20. At about ρ = 0.80, MR begins to significantly outperform mixed model. **Table 2**: Comparison of 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. Correlations among predictors are as high as 0.90, so we expect the addition of the ridge component to improve prediction ability. **Figure 2**: Normal QQ plot of t-statistics for MR. Points circled in red are significant for MR, while the point circled in gray is significant for the mixed model.

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 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 APOb as significantly associated with BP over time while the usual mixed modeling approach was unable to detect this association.

ACKNOWLEDGEMENTS

Support for this research was provided by NIH award R01HL107196.