Testing for the Effect of a Genetic Pathway in Longitudinal/Clustered Data

with Application to DNA Methylation Data

Arnab Maity

Department of Statistics
North Carolina State University

Co-authors: Stacey E. Alexeeff and Xihong Lin

Outline

1. Introduction and Motivation
2. Kernel Machine Based Approach
3. Simulation and Data Results

Outline

1. Introduction and Motivation
2. Kernel Machine Based Approach
3. Simulation and Data Results

Longitudinal Data: Normative Aging Study (NAS)

Objective

Test association between the methylation of pro-inflammatory genes and cardiovascular biomarkers.

NAS is a ongoing cohort study of older men in Boston, MA.

- 277 men ages 50-100 living in the greater Boston area
- Physical evaluations every 3 years
- Measurements of gene specific methylation at 2 visits

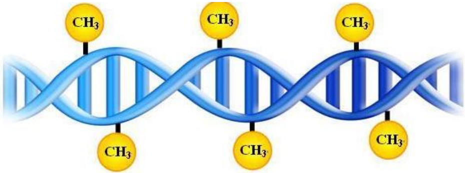
Outcome: C-reactive protein (CRP)

- Biomarker of inflammation and is predictive of cardiovascular events

DNA Methylation

DNA Methylation

- Addition of methyl groups to CpG sites in the DNA
- Modifies genome function without changing the DNA sequence



DNA Methylation

- An epigenetic regulator of gene expression.
- Methylation status or patterns are involved in determining cardiovascular diseases (CVD).
- Modifiable in response to environmental factors.
- De-methylation of DNA in a particular gene is expected to increase the expression and activity of the protein coded by the gene.

Genes of interest

- Toll-like receptor 2 (TLR2)
- Interferon gamma (IFN-γ)
- Intercellular adhesion molecule 1 (ICAM-1)
- Interleukin 6 (IL-6)

Scientific Questions

Outcome: C-reactive protein (CRP)

1. Are Toll-like receptor 2 (TLR2) and Interferon gamma (IFN-γ) associated with CRP levels?
    - Function: TLR2 regulates IFN-γ

2. Are Intercellular adhesion molecule 1 (ICAM-1) and Interleukin 6 (IL-6) associated with CRP levels?
    - Function: on same inflammatory pathway

Objective

- Scientific goal: test for genetic effects on disease outcome
  - A set of genetic covariates may be on the same biological pathway
  - Genes within a pathway may be correlated, and may interact in functional ways
  - Outcome/genes between visits are correlated

Outline

1. Introduction and Motivation
2. **Kernel Machine Based Approach**
3. Simulation and Data Results

Longitudinal Modeling Framework

For each subject $i = 1, \ldots, n$ at time $j = 1, \ldots, J$,

- Y_{ij} is a continuous response
- $\mathbf{Z}_{ij} = (Z_{ij1}, \ldots, Z_{ijM})^T$ are M genetic covariates
- $\mathbf{X}_{ij}$ is a set of clinical covariates

We assume the following model:

$$Y_{ij} = \mathbf{X}_{ij}^T \beta + h(\mathbf{Z}_{ij}) + \epsilon_{ij}$$

where,

- β is a unknown parameter vector
- $h(\cdot)$ is an unknown function
- $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{iJ}) = Normal\{0, R(\tau)\}$

Framework

Test for $H_0 : h(\mathbf{z}) = h(z_1, \ldots, z_M) = 0$

- This formulation covers a broad range of models
  - Main effects only model:

$$h(z_1, \ldots, z_M) = z_1\eta_1 + \ldots + z_M\eta_M$$

$$H_0 : \eta_1 = \ldots = \eta_M = 0$$

  - First-order interaction model:

$$h(z_1, \ldots, z_M) = \sum_{j=1}^{M} z_j\eta_j + \sum_{j<k} z_j z_k \gamma_{jk}$$

$$H_0 : \eta_1 = \ldots = \eta_M = \gamma_{12} = \ldots = \gamma_{M-1,M} = 0$$

  - Nonparametric formulation of $h(\cdot)$ is also allowed

Parametric Problem

- Assume that $h(\cdot)$ has a parametric form, e.g., the first order interaction model
- Usual test for H_0 is the F-test with $M(M+1)/2$ degrees of freedom

Drawbacks of the usual F-test

- Uses large degrees of freedom resulting in power loss
- The parametric assumption on $h(\cdot)$ might be too strong and not flexible
- Power loss due to possible correlation among the genes or correlation among the visits

Kernel Machine Formulation

- Assume the function $h : R^M \to R$ resides in a function space $\mathcal{H}_K$ with a positive semidefinite reproducing kernel $K : R^M \times R^M \to R$

- A kernel function $K(\mathbf{z}, \mathbf{z}')$ has two arguments
  - $\mathbf{z}$: covariate vector for subject 1
  - $\mathbf{z}'$: covariate vector for subject 2

- $K(\mathbf{z}, \mathbf{z}')$ measures the 'similarity' (or 'dissimilarity') between these covariate vectors

- By Riesz representation theorem

$$h(\mathbf{z}) = \langle h, K(\mathbf{z}, \cdot) \rangle_{\mathcal{H}_K}$$

Kernel Machine Formulation

- Two ways to characterize h
  - Using basis functions (primal form): corresponds to regular regression

$$h(\mathbf{z}) = \sum_{\ell=1}^{L} \phi_\ell(\mathbf{z})\eta_\ell$$

  - Using a positive definite kernel function $K(\cdot,\cdot)$ (dual form):

$$h(\mathbf{z}) = \sum_{i,j} K(\mathbf{Z}_{ij}, \mathbf{z})\alpha_{ij}$$

Mercer's theorem (Cristianini and Shawe-Taylor, 2000)

The kernel function $K(\cdot,\cdot)$ implicitly specifies a unique function space spanned by a particular set of orthogonal basis functions.

Kernel Machine Formulation

Examples

- Linear kernel: $K(\mathbf{z}, \mathbf{z}') = 1 + z_1 z_1' + \ldots + z_M z_M'$
  - Basis representation: $\phi(\mathbf{z}) = [z_1, \ldots, z_M]$
- Quadratic kernel: $K(\mathbf{z}, \mathbf{z}') = (1 + z_1 z_1' + \ldots + z_M z_M')^2$
  - Basis representation: $\phi(\mathbf{z}) = [z_1, \ldots, z_M, z_1^2, \ldots, z_M^2, z_1 z_2, \ldots, z_{M-1} z_M]$
- Gaussian kernel: $K(\mathbf{z}, \mathbf{z}') = \exp\{-\sum_{j=1}^{M} (z_j - z_j')^2 / \delta\}$
  - Basis representation: space spanned by radial basis

Methodology

- Model: $Y_{ij} = \mathbf{X}_{ij}^T \beta + h(\mathbf{Z}_{ij}) + \epsilon_{ij}$
- Penalized log-likelihood:

$$-\sum_{i,j,k} \{Y_{ij} - \mathbf{X}_{ij}^T \beta - h(\mathbf{Z}_{ij})\} R^{jk}(\tau) \{Y_{ik} - \mathbf{X}_{ik}^T \beta - h(\mathbf{Z}_{ik})\} - \lambda^{-1} ||h||^2$$

- Kernel representation: $h(\mathbf{z}) = \sum_{i=1}^n \sum_{j=1}^J K(\mathbf{Z}_{ij}, \mathbf{z}) \alpha_{ij} = \mathbf{K}\alpha$

Kernel log-likelihood

$$-\{\mathbf{Y} - \mathbf{X}^T \beta - \mathbf{K}^T \alpha\} R^{-1}(\tau) \{\mathbf{Y} - \mathbf{X}^T \beta - \mathbf{K}^T \alpha\} - \lambda^{-1} \alpha^T \mathbf{K} \alpha$$

Normal Equations

$$\begin{bmatrix} \mathbf{X}^{\mathrm{T}}\widetilde{R}^{-1}\mathbf{X} & \mathbf{X}^{\mathrm{T}}\widetilde{R}^{-1}\mathbf{K} \\ \widetilde{R}^{-1}\mathbf{X} & \lambda\mathbf{K}^{-1}+\widetilde{R}^{-1} \end{bmatrix} \begin{pmatrix} \beta \\ \mathbf{h} \end{pmatrix} = \begin{bmatrix} \mathbf{X}^{\mathrm{T}}\widetilde{R}^{-1}\mathbf{Y} \\ \widetilde{R}^{-1}\mathbf{Y} \end{bmatrix},$$

where $\widetilde{R} = I_n \otimes R$

Mixed Model Formulation

- Penalized formulation is equivalent to mixed model

$$Y_{ij} = \mathbf{X}_{ij}^T \beta + h_{ij} + \epsilon_{ij},$$

where

$$\begin{aligned} \mathbf{h} &= (h_{11}, \ldots, h_{nJ})^T = Normal(0, \lambda K) \\ \epsilon &= (\epsilon_{11}, \ldots, \epsilon_{nJ})^T = Normal(0, I_n \otimes R(\tau)) \end{aligned}$$

- $\widehat{\beta}$: Best Linear Unbiased Estimator (BLUE)
- $\widehat{\mathbf{h}}$: Best Linear Unbiased Predictor (BLUP)

Testing

Hypothesis

$$H_0 : h(\cdot) = 0 \iff H_0 : \lambda = 0$$

- Restricted log-likelihood

$$\begin{aligned} L_{\mathrm{REML}} &= -\log|\mathbf{V}|/2 - \log|\mathbf{X}\mathbf{V}^{-1}\mathbf{X}^T|/2 \\ &\quad -(\mathbf{Y} - \mathbf{X}^T\beta)^T\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}^T\beta)/2, \end{aligned}$$

  where $\mathbf{V} = I_n \otimes R(\tau) + \lambda\mathbf{K}$.
- Score test
  - Score statistic is a quadratic form
  - Null distribution can be computed

Outline

1 Introduction and Motivation

2 Kernel Machine Based Approach

3 Simulation and Data Results

Simulation

- Number of simulations: 2,000
- $n = 100$ subjects, $J = 3$ time points per subject
- Health outcome Y_{ij} depends linearly on X_{ij} and $\mathbf{Z}_{ij}$,

$$Y_{ij} = \mathbf{X}_{ij}^T \beta + \mathbf{Z}_{ij}^T \gamma + \epsilon_{ij}$$

- 3 clinical covariates $\mathbf{X}_{ij}$ simulated with some correlation within subject
- 10 genetic covariates $\mathbf{Z}_{ij}$ were continuous and simulated to induce correlation within subject-visit and within a given gene across visits
- $\beta = (0.7, 0.7, 0.7)$ and $\gamma = (c, \ldots, c)$
- ϵ_{ij} simulated with compound symmetry structure, where several values of correlation (ρ) were considered: $0, 0.03, 0.06$

Simulation

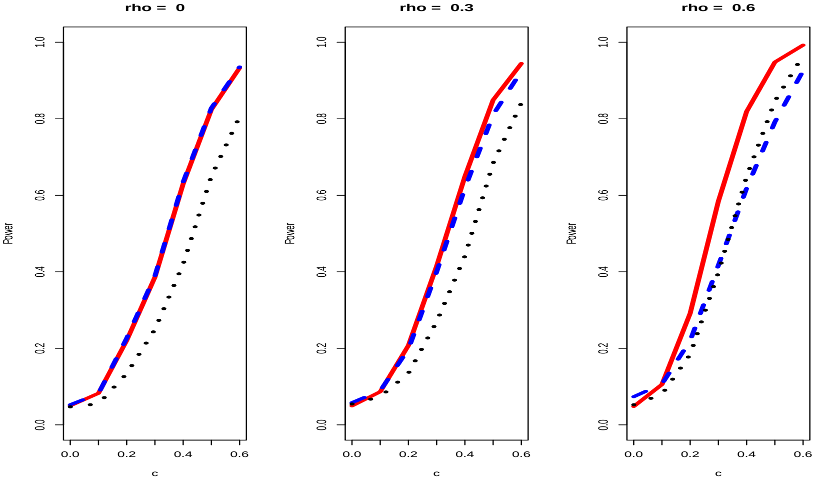
For each set of simulations,

- Size: $c = 0$ (no dependence on Z)
- Power: $c = 0.1, 0.2, \ldots, 0.6$

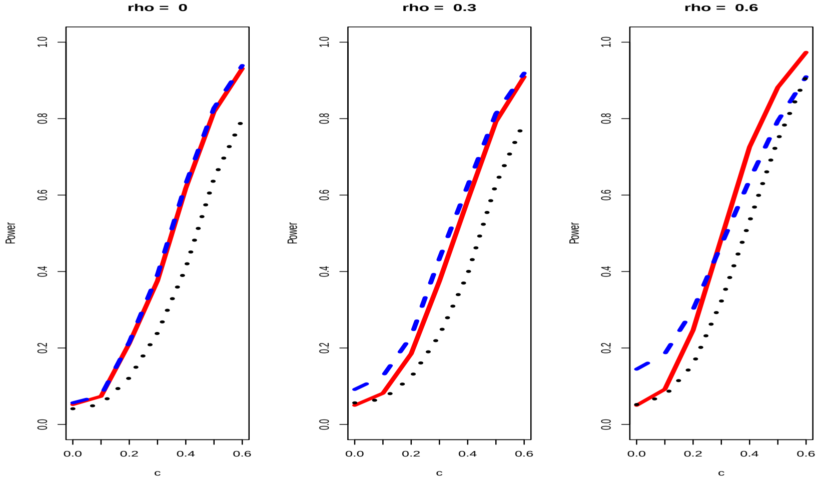
We considered 3 simulation cases with different correlations in the genes:

1. $Corr(Z_{ijm}, Z_{ij'm}) = 0.17$, $Corr(Z_{ijm}, Z_{ijm'}) = 0.17$
2. $Corr(Z_{ijm}, Z_{ij'm}) = 0.4$, $Corr(Z_{ijm}, Z_{ijm'}) = 0.4$
3. $Corr(Z_{ijm}, Z_{ij'm}) = 0.75$, $Corr(Z_{ijm}, Z_{ijm'}) = 0.33$

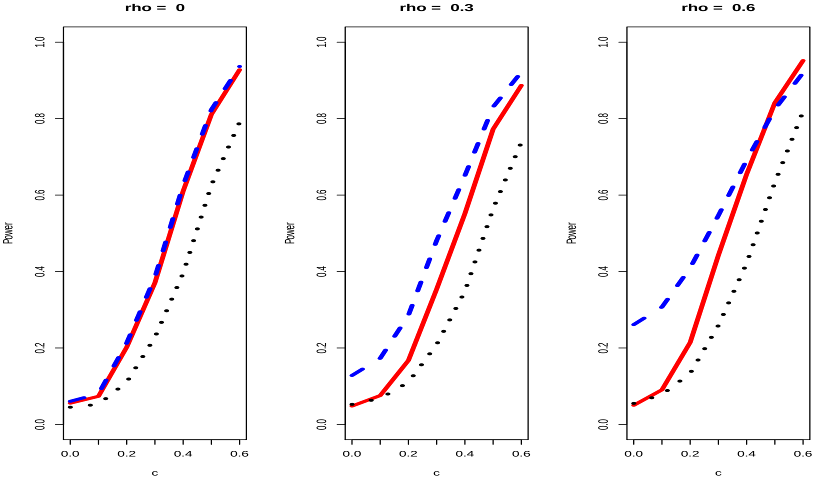
Results for Case 1: Low correlation



Results for Case 2: Moderate correlation



Results for Case 3: High correlation



Longitudinal Data: Normative Aging Study (NAS)

Objective

Test association between the methylation of pro-inflammatory genes and cardiovascular biomarkers.

- $n = 277$ men ages 50-100 living in the greater Boston area
- Outcome: C-reactive protein (CRP)
- Measurements of gene specific methylation at 2 visits

Gene-set of interest

- Toll-like receptor 2 (TLR2) and Interferon gamma (IFN-γ)
- Intercellular adhesion molecule 1 (ICAM-1) and Interleukin 6 (IL-6)

NAS Data Analysis

Toll-like receptor 2 (TLR2) and Interferon gamma (IFN-γ)

- Function: TLR2 regulates IFN-γ
- TLR2 measured at 5 positions
- IFN-γ measured at 2 positions

Results

- F test: 7 df, p-value= 0.005
- Independent KM test: 1.972 df, p-value= 0.00000932
- Longitudinal KM test: 2.047 df, p-value= 0.00001937 (corr = 0.17 in residual errors)

NAS Data Analysis

Intercellular adhesion molecule 1 (ICAM-1) and Interleukin 6 (IL-6)

- Function: on same inflammatory pathway
- ICAM-1 measured at 3 positions
- IL-6 measured at 2 position

Results

- F test: 5 df, p-value= 0.1411
- Independent KM test: 1.29 df, p-value= 0.0849
- Longitudinal KM test: 1.35 df, p-value= 0.0610 (corr = 0.23 in residual errors)

Summary

- In simulations, we see better performance when accounting for correlation in both the genetic data and in the residual errors
- Preliminary analysis of data suggests methylation of pro-inflammatory genes may be associated with CRP levels.

Next steps:

- Testing for individual genes
- Different correlation structures
- More in-depth analysis of data