Statistics and Probability Letters 94 (2014) 248–256

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Assessing the relationship of evolutionary rates and functional variables by mixture estimating equations

Pei-Sheng Lin [a,b,*], Feng-Chi Chen [a,c], Shu-Fu Kuo [a], Yi-Hung Kung [a]

[a] Division of Biostatistics and Bioinformatics, National Health Research Institutes, Taiwan

[b] Department of Mathematics, National Chung Cheng University, Taiwan

[c] Department of Dentistry, China Medical University, Taiwan

ARTICLE INFO

Article history:
Received 18 July 2013
Received in revised form 11 June 2014
Accepted 29 July 2014
Available online 8 August 2014

Keywords:
Evolutionary rate
Generalized estimating equation
Mixture distribution

ABSTRACT

In the study of complex organisms, clarifying the association between the evolution of coding genes and the measures of functional variables is of fundamental importance. However, traditional analysis of the evolutionary rate is either built on the assumption of independence between responses or fails to handle a mixture distribution problem. In this paper, we utilize the concept of generalized estimating equations to propose an estimating equation to accommodate continuous and binary probability distributions. The proposed estimate can be shown to have consistency and asymptotic normality. Simulations and data analysis are also presented to illustrate the proposed method.

1. Introduction

In complex organisms, important biological features are usually influenced by multiple factors. One good example is the evolution of coding genes. Clarifying the determinants of coding sequence evolution is of fundamental importance to research on genome evolution and molecular biology. However, one complexity of such biological studies is the existence of "subunits". For instance, the expression levels and patterns should be similar between complete genes and their component exons, but may differ by log scales between different genes (and their exons). In other words, when dealing with this biological property (expression level), we can group "exons" into larger units (i.e., "genes"). One additional layer of complexity in such analyses comes from the inter-correlations between the involved factors. Examples include the association between subcellular localization of a protein and its topological feature in the protein interaction network. New methodologies are thus required to address this high level of complexity in biological analysis.

The purpose of this paper is to develop a statistical method to find significant associations between the evolution of coding exons (such as d_N; nonsynonymous nucleotide substitution rate) and the measures of functional variables. The properties of the data add complexity. First, the dependent variable d_N is correlated with one another for exons of the same gene. Second, some of the biological features are also correlated with one another for exons of the same gene. Third, the analysis includes both categorical and continuous variables, with the latter varying by several log scales. One traditional approach to analyzing such data is first to use principal component analysis (PCA) to group the functional variables, and then to use regression analysis to model the evolutionary rate by grouped variables. This estimation process is referred to as a PCR analysis (e.g. Chen et al., 2012). However, we note that PCR analysis is built on an independence assumption for the responses. Because the (observed) evolutionary rates among the same gene may have their own correlation structure

* Correspondence to: Division of Biostatistics and Bioinformatics, National Health Research Institutes, 35 Keyan Road, Zhunan, Miaoli County 350, Taiwan.
E-mail addresses: pslin@nhri.org.tw, pslin@math.ccu.edu.tw (P.-S. Lin).

http://dx.doi.org/10.1016/j.spl.2014.07.031
0167-7152/

different from that of functional variables, PCR analysis of this may lead to misleading statistical inference. Another problem is that PCR analysis may have difficulty dealing with non-normal data, particularly when the data set has a certain proportion of zero values. Other approaches to systematically examining the contributions of these inter-correlated factors to the evolutionary rate – including multiple regression (Jovelin and Phillips, 2009), partial correlation (Liao et al., 2006) and probabilistic modeling (Xia et al., 2009) – have similar limitations.

On the other hand, the generalized estimating equation (GEE), which was developed for longitudinal data, can be used to analyze the relationship between the evolutionary rate and functional variables. Since the evolutionary rates are correlated within a specific gene but are independent of other genes, such a data set with block correlations can suitably fit the GEE assumption. However, the traditional GEE is to assume that the responses are from the same exponential family distribution. When analyzing the evolutionary data, we find that the evolutionary rate, such as d_N, may follow a mixture distribution function since some of the responses are zeros and the others are positive values in an interval. As mentioned by Olsen and Schafer (2001), using the traditional generalized estimating equation would be infeasible to recognize the qualitative distinctions between zero and non-zero responses.

To address the above issues, we modified the GEE concept to accommodate two probability functions via a latent process. In Section 2, we introduce the mixture model and how to construct an estimating equation based on the concept of GEEs for this model. To select the variables, we rely on the quasi-deviance (QDEV) function by Lin (2011). A simulation study for the proposed method is given in Section 3. In Section 4, we use a data set concerning the difference of evolutionary rates between human and mouse to illustrate the proposed method. Some discussions are presented in Section 5.

2. Estimation methods

2.1. Standard generalized estimating equations

Let $Y_{i,j}$ be the jth observation in the ith cluster with the expectation $\theta_{i,j} = E(Y_{i,j})$, $i = 1, \ldots, K$, $j = 1, \ldots, n_i$. In the evolutionary data, a single gene with n_i observed exons is regarded as a cluster; thus, the total number of observations is $N = \sum_{i=1}^{K} n_i$. Let $\boldsymbol{x}_{i,j} = (x_{i,j,0}, x_{i,j,1}, \ldots, x_{i,j,q})^t$ be a vector of explanatory variables associated with $Y_{i,j}$, where $x_{i,j,0} \equiv 0$ and notation t denotes the transpose of a matrix. Assume that the marginal distribution $F_{i,j}$ of $Y_{i,j}$ belongs to an exponential family distribution. To fit a model for $Y_{i,j}$ by $\boldsymbol{x}_{i,j}$, we consider a generalized linear model $\theta_{i,j} = g(\boldsymbol{x}_{i,j}^t\boldsymbol{\beta})$, where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_q)'$ is a R^{q+1} vector of regression parameters associated with explanatory variables and $g(\cdot)$ is a link function.

We now define some notations. Let $\boldsymbol{Y}_i = (Y_{i,1}, \ldots, Y_{i,n_i})^t$ and $\boldsymbol{Y} = (\boldsymbol{Y}_1^t, \ldots, \boldsymbol{Y}_K^t)^t$ denote the responses in the ith cluster and the whole data, respectively. Also, we define $\boldsymbol{\theta}_i = E(\boldsymbol{Y}_i)$ and $\boldsymbol{V}_i = \mathrm{var}(\boldsymbol{Y}_i)$ as the expectation and variance–covariance matrices of $\boldsymbol{Y}_i$, respectively. Let $\mathrm{corr}(Y_{i,j}, Y_{i',j'})$ be the correlation between $Y_{i,j}$ and $Y_{i',j'}$. In the cluster data, we assume that $\mathrm{corr}(Y_{i,j}, Y_{i',j'}) = \rho_{j,j'}^{i}$ for $i = i'$ and $\mathrm{corr}(Y_{i,j}, Y_{i',j'}) = 0$ for $i \neq i'$. Let $\boldsymbol{V} = \mathrm{var}(\boldsymbol{Y})$. Then, in the cluster data, we have $\boldsymbol{V} = \mathrm{diag}(\boldsymbol{V}_1, \ldots, \boldsymbol{V}_K)$, where diag denotes a block-diagonal matrix with diagonal elements $\boldsymbol{V}_1, \ldots, \boldsymbol{V}_K$.

To analyze the cluster data, a useful tool is to use the generalized estimating equation (Liang and Zeger, 1986)

$$\sum_{i=1}^{K} \boldsymbol{D}_i^t \boldsymbol{V}_i^{-1}\{\boldsymbol{Y}_i - \boldsymbol{\theta}_i\}\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_g} = \boldsymbol{0}, \tag{2.1}$$

where $\boldsymbol{D}_i = \partial\boldsymbol{\theta}_i/\partial\boldsymbol{\beta}$ is a derivative matrix of $\boldsymbol{\theta}_i$ with respect to $\boldsymbol{\beta}$. The estimate $\hat{\boldsymbol{\beta}}_g$ is called a GEE estimate. Let $\boldsymbol{S}_i(\boldsymbol{\beta}) = \boldsymbol{D}_i^t\boldsymbol{V}_i^{-1}\{\boldsymbol{Y}_i - \boldsymbol{\theta}_i(\boldsymbol{\beta})\}$ and $\boldsymbol{I}_0(\boldsymbol{\beta}) = \sum_{i=1}^{K} \boldsymbol{D}_i^t\boldsymbol{V}_i^{-1}\boldsymbol{D}_i$. When the link function $g(\cdot)$ is non-linear, we usually use a Newton–Raphson iteration, $\hat{\boldsymbol{\beta}}_g^{(l+1)} = \hat{\boldsymbol{\beta}}_g^{(l)} + \boldsymbol{I}_0^{-1}\{\hat{\boldsymbol{\beta}}_g^{(l)}\}\boldsymbol{S}_i\{\hat{\boldsymbol{\beta}}_g^{(l)}\}$, to get an approximation for $\hat{\boldsymbol{\beta}}_g$, where (l) denotes the lth iteration.

2.2. Generalized estimating equations for mixture distributions

In practice, the evolutionary rate may have a large proportion of zero values. That is, the distribution function can be written as $dF_{i,j} = \{P(Y_{i,j} = 0)\}^{I[y_{i,j}=0]}\left[\{1 - P(Y_{i,j} = 0)\}f(y_{i,j})\right]^{I[y_{i,j}>0]} d\tau d\mu$, where $f(\cdot)$ is a probability density function of continuous random variables, and μ and τ are Lebesgue and counting measures, respectively. Since the distribution function $F_{i,j}$ does not belong to the exponential family distribution, the traditional GEE framework (2.1) may not work for the mixture case.

In addition, in analysis of the evolutionary data, we often find that a positive value of $Y_{i,j}$ seems to follow a normal distribution after taking a log-transformation. To fit a model satisfying this pattern and a mixture distribution, we assume that a latent Gaussian process $\epsilon_{i,j}$ with mean zero, variance σ^2, and correlation $\rho_{j,j'}^{i}$ exists. Given $\epsilon_{i,j}$, the response $Y_{i,j}$ is independently generated by

$$Y_{i,j} = \begin{cases} 0, & \text{if } h(\boldsymbol{x}_{i,j}^t\boldsymbol{\beta} + \epsilon_{i,j}) \geq k_0, \\ \exp(\boldsymbol{x}_{i,j}^t\boldsymbol{\beta} + \epsilon_{i,j}), & \text{if } h(\boldsymbol{x}_{i,j}^t\boldsymbol{\beta} + \epsilon_{i,j}) < k_0, \end{cases} \tag{2.2}$$