Journal of Multivariate Analysis 147 (2016) 202–217

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Simultaneous variable selection and de-coarsening in multi-path change-point models

Azadeh Shohoudi, Abbas Khalili*, David B. Wolfson, Masoud Asgharian

Department of Mathematics and Statistics, McGill University, Montréal, Québec, Canada H3A 0B9

ARTICLE INFO

Article history:
Received 19 November 2014
Available online 10 February 2016

AMS subject classifications:
primary 62J02
62F10
62J07
secondary 62F12

Keywords:
Change-point models
EM algorithm
Regularization
LASSO
SCAD
Alzheimer's disease

ABSTRACT

Follow-up studies on a group of units are commonly carried out to explore the possibility that a response distribution has changed at unobservable time points that are different for different units. Often, in practice, there will be many potential covariates, which may not only be associated with the response distribution but also with the distribution of the unobservable change-points. Here, the covariates are allowed to enter the change-point distribution through a proportional odds model whose baseline odds is assumed to be piecewise constant as a function of time. The combination of a large number of putative regression coefficients in the response distributions as well as the change-point distribution, alone leads to a challenging simultaneous variable selection and estimation problem. Moreover, selection and estimation of the parameters that determine the coarseness of the baseline odds function adds a further level of complexity. Using penalized likelihood methods we are able to simultaneously perform variable selection, estimation, and determine the coarseness of the baseline odds function. Our approach is computationally efficient and shown to be consistent in variable selection and parameter estimation. We assess its performance through simulations, and demonstrate its usage in fitting a model for cognitive decline in subjects with Alzheimer's disease.

1. Introduction

In Alzheimer's disease (AD), the rate of progression is highly variable and there has been much interest in the identification of factors associated with cognitive decline [6,10]. Similarly, in clinical trials interest may be in factors associated with the expected responses following the administration of a treatment [4]. Further, floor and ceiling effects in tests of cognition such as the Mini-Mental State Exam (MMSE), and a delay to treatment effect in clinical trials, induce changes in the response distribution at time points that are not directly observable. In this multi-subject (i.e., multi-path) change-point setting, it is expected that the response distribution as well as the distribution of the change-points will be a function of subject-specific covariates.

Often, the number of covariates initially entertained is large although many may be spurious and must be selected out. The combination of a large number of initial regression coefficients in the response distributions as well as the change-point distribution, alone leads to a challenging simultaneous variable selection and estimation problem. We allow the covariates to enter the change-point distribution through a proportional odds model whose baseline odds is assumed to be piecewise constant as a function of time. This results in an additional level of complexity where it is desired to carry out simultaneous selection and estimation of the parameters that determine the coarseness of the baseline odds function. For instance,

* Corresponding author.
E-mail address: khalili@math.mcgill.ca (A. Khalili).

http://dx.doi.org/10.1016/j.jmva.2016.02.001

in the illustrative Alzheimer's disease application that we present in Section 7, even with a small number of covariates, simultaneous de-coarsening and variable selection leads to a forbidding number, 2^{46}, of submodels. We overcome the difficulties that frequently accompany variable selection and estimation in multi-path change-point problems by developing a methodology based on a penalized likelihood approach.

Multi-path change-point models with covariates in the change-point distribution, as well as the response distributions, were presented in [2]. The discussion in this paper did not, however, address the variable selection problem and was limited to maximum likelihood estimation.

Most of the research on variable selection using familiar penalty functions such as the LASSO [19], SCAD [8], adaptive LASSO [25], fused LASSO [20], and the smooth integration of counting and absolute deviation (SICA, [15]), has focused on linear and generalized linear regression models. More recently, attention has turned to segmented regression models where the cut-points are unknown. In particular, this research has focused on a single-path change-point model with the assumption that the responses are independent random variables. For example, Wu [23] proposed an information criterion for simultaneous change-point detection and variable selection in a linear regression model with a possible change-point. Harchaoui and Lévy-Leduc [11] proposed a method based on the LASSO/LARS for estimating the locations of multiple change-points in a single-path piecewise constant regression function. Ciuperca [5] used the LASSO for variable selection and estimation in a single-path change-point model with a fixed number of regression segments. She extended her methods to an unknown number of segments. As will be seen in Section 2, the multi-path change-point model that we discuss here is quite different from those considered in the above papers, and to the best of our knowledge the problem of variable selection as well as de-coarsening in this model has not been addressed. Such scenarios are particularly important for analyzing multi-subject longitudinal data that are frequently collected in the medical and other fields.

The layout of the paper is as follows. In Section 2, we formally introduce the multi-path change-point model with covariates. In Section 3, we present our penalized likelihood approach to the problems of variable selection, estimation, and de-coarsening. In Section 4, we describe the algorithm for numerical computations. Section 5 contains the asymptotic properties of our method. An investigation of its finite-sample properties through simulations is described in Section 6. We demonstrate the usage of the proposed method when fitting a model for cognitive decline in subjects with Alzheimer's disease in Section 7. Section 8 contains closing remarks.

2. Terminology and the model

Let observations on a scalar-valued response variable $Y$and a covariate-vector $\boldsymbol{X} = (X_1, \dots, X_p)^\top$ be taken on the time interval $[0, T]$, for n subjects at equally spaced time points, $0 = t_1 < \cdots < t_m = T$.

For $i \in \{1, \dots, n\}$, let $(\boldsymbol{Y}_i, \boldsymbol{X}_i) = (Y_{i1}, \dots, Y_{im}, X_{i1}, \dots, X_{ip})$ denote the random vector of observations for subject i. We denote the corresponding realized values by $(\boldsymbol{y}_i, \boldsymbol{x}_i) = (y_{i1}, \dots, y_{im}, x_{i1}, \dots, x_{ip})$. We shall call a discrete random variable $\tau_i < m$ a change-point, if conditional on $\tau_i = k$ and the covariate values $\boldsymbol{x}_i$, the joint distribution of the responses before and including k is different from the joint distribution of the responses after k. The τ_i's are not directly observable and can be considered as latent variables.

Let $f_1^*(y_{i\ell}; \boldsymbol{\theta}_1(\boldsymbol{x}_i), \phi_1)$ and $f_2^*(y_{i\ell}; \boldsymbol{\theta}_2(\boldsymbol{x}_i), \phi_2)$ be the respective marginal conditional probability density functions before and after the change. We assume that conditional on $\tau_i = k$ and $\boldsymbol{x}_i$, the responses are independent. For a possible relaxation of this assumption see Section 8. Under conditional independence we have

$$(Y_{i1}, \dots, Y_{ik})\big|\boldsymbol{x}_i \sim \prod_{\ell=1}^{k} f_1^*(y_{i\ell}; \boldsymbol{\theta}_1(\boldsymbol{x}_i), \phi_1),$$

$$(Y_{i,k+1}, \dots, Y_{im})\big|\boldsymbol{x}_i \sim \prod_{\ell=k+1}^{m} f_2^*(y_{i\ell}; \boldsymbol{\theta}_2(\boldsymbol{x}_i), \phi_2)$$

for $k \in \{1, \dots, m-1\}$. Consequently, the conditional joint density function for the responses of the ith subject, given $\tau_i = k$ and $\boldsymbol{x}_i$, is

$$f_k(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\Upsilon}) = \prod_{\ell=1}^{k} f_1^*(y_{i\ell}; \boldsymbol{\theta}_1(\boldsymbol{x}_i), \phi_1) \prod_{\ell=k+1}^{m} f_2^*(y_{i\ell}; \boldsymbol{\theta}_2(\boldsymbol{x}_i), \phi_2). \tag{1}$$

If $k = m$, by convention, no change is said to have occurred and

$$f_m(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\Upsilon}) = \prod_{\ell=1}^{m} f_1^*(y_{i\ell}; \boldsymbol{\theta}_1(\boldsymbol{x}_i), \phi_1).$$

We assume $\boldsymbol{\theta}_j(\boldsymbol{x}_i) = g(\beta_{j0} + \boldsymbol{x}_i^\top \boldsymbol{\beta}_j)$ for $j = 1, 2$, where g is a known link function and $(\beta_{10}, \boldsymbol{\beta}_1) = (\beta_{10}, \beta_{11}, \dots, \beta_{1p})^\top$ and $(\beta_{20}, \boldsymbol{\beta}_2) = (\beta_{20}, \beta_{21}, \dots, \beta_{2p})^\top$ are vectors of regression parameters before and after the change. We denote the dispersion parameters in the response distributions before and after the change by ϕ_1 and ϕ_2. The vector of all parameters of the response distributions is denoted by $\boldsymbol{\Upsilon} = (\beta_{10}, \boldsymbol{\beta}_1, \beta_{20}, \boldsymbol{\beta}_2, \phi_1, \phi_2)$.