Computational Statistics and Data Analysis 95 (2016) 1–16

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Bayesian variable selection for finite mixture model of linear regressions

Kuo-Jung Lee [a], Ray-Bing Chen [a,*], Ying Nian Wu [b]

[a] Department of Statistics, National Cheng Kung University, Taiwan

[b] Department of Statistics, University of California, Los Angeles, United States

ARTICLE INFO

Article history:
Received 16 July 2014
Received in revised form 11 September 2015
Accepted 12 September 2015
Available online 9 October 2015

Keywords:
Gibbs sampler
Median probability criterion
Sparsity
Stochastic search variable selection

ABSTRACT

We propose a Bayesian variable selection method for fitting the finite mixture model of linear regressions. The model assumes that the observations come from a heterogeneous population which is a mixture of a finite number of sub-populations. Within each sub-population, the response variable can be explained by a linear regression on the predictor variables. If the number of predictor variables is large, it is assumed that only a small subset of variables are important for explaining the response variable. It is further assumed that for different sub-populations, different subsets of variables may be needed to explain the response variable. This gives rise to a complex variable selection problem. We propose to solve this problem within the Bayesian framework where we introduce two sets of latent variables. The first set of latent variables are membership indicators of the observations, indicating which sub-population each observation comes from. The second set of latent variables are inclusion/exclusion indicators for the predictor variables, indicating whether or not a variable is included in the regression model of a sub-population. Variable selection can then be accomplished by sampling from the posterior distributions of the indicators as well as the coefficients of the selected variables. We conduct simulation studies to demonstrate that the proposed method performs well in comparison with existing methods. We also analyze a real data set to further illustrate the usefulness of the proposed method.

1. Introduction

Variable selection is a fundamental problem in linear regression and has become increasingly important for many modern applications. During the past decade, a rich literature has been developed around this problem, especially for the case where large numbers of variables are collected and the number of variables exceeds the number of observations. The methods proposed for the problem of variable selection can be roughly classified into two categories.

One category consists of various penalized least squares methods, including the famous Lasso method (Tibshirani, 1996) based on the convex ℓ_1 penalty for regularization, as well as non-convex penalties such as SCAD (Fan and Li, 2001) and MCP (Zhang, 2010). The Lasso approach has also been extended to more sophisticated forms such as the group Lasso and graphical Lasso; see (Tibshirani, 2011) for a review.

The other category consists of various Bayesian variable selection methods, such as stochastic search variable selection (SSVS) (George and McCulloch, 1993; Chen, 2012) and Bayesian Lasso (Park and Casella, 2008). The Bayesian method assumes

* Corresponding author.
E-mail addresses: kjlee@stat.ncku.edu.tw (K.-J. Lee), rbchen@stat.ncku.edu.tw (R.-B. Chen), ywu@stat.ucla.edu (Y.N. Wu).

http://dx.doi.org/10.1016/j.csda.2015.09.005
0167-9473/

a prior distribution on the regression coefficients. A popular prior is the so-called "spike and slab" prior, which is a mixture of a point mass at zero and a diffused Gaussian distribution. Such a prior distribution assumes a binary indicator for each variable, indicating whether a variable is included in the regression or not. Variable selection can be accomplished by sampling from the posterior distributions of the latent indicators and the posterior distributions of the coefficients of the selected variables.

The two categories of methods are related in that the penalty terms correspond to specific Bayesian prior distributions. The penalized least squares approaches, especially the Lasso and its extensions, usually enjoy a computational advantage since the objective functions are convex and can be easily minimized.

Despite the wide applicability of the linear regression model powered by modern variable selection tools, a single regression model can be inadequate if the data come from a heterogeneous population that consists of a number of different sub-populations with different characteristics. In this situation, it is possible that a separate linear regression model is needed for each sub-population. Moreover, the regression models in different sub-populations may use different subsets of predictor variables (or regressors, covariates) to explain the response variable. If the memberships of the observations are unobserved, then we naturally have a finite mixture model of linear regressions, where each mixture component is a linear regression model with its own subset of predictor variables. This gives rise to a variable selection problem that is more complex than that of a single linear regression model.

To solve the variable selection problem in the mixture model, one may extend the penalized least squares methods to the penalized likelihood of mixture models (Khalili and Chen, 2007; Städler et al., 2010). However, the negative log-likelihood of the mixture model would no longer be convex, causing it to lose one of the most appealing features of the Lasso method and its various extensions. As a result, one can only draw inference based on local minima of the objective functions. This difficulty motivates us to adopt a Bayesian alternative which appears more natural.

Specifically, the Bayesian variable selection method for the mixture model involves two sets of latent variables or indicators. The first set of latent variables are membership indicators associated with the observations, indicating which sub-population each observation comes from. The second set of latent variables are inclusion/exclusion indicators for the variables, where, for each sub-population or mixture component, a binary indicator is associated with each predicator variable, indicating whether or not this variable is included in the linear regression of this mixture component. Variable selection, clustering, and parameter estimation can then be carried out by sampling from the posterior distributions of the indicators and the posterior distributions of the coefficients of the selected variables. Our simulation studies show that the Bayesian method performs well compared with existing methods and that the corresponding Markov chain Monte Carlo (MCMC) algorithm may be capable of escaping the traps of local minima. We also analyze a real data set to further illustrate the usefulness of the proposed method.

The rest of the article is organized as follows. Section 2 presents the finite mixture model of linear regressions and its Bayesian treatment, including the prior specifications and the MCMC algorithm for posterior sampling. Section 3 illustrates our method by simulation studies, where our method is compared with existing methods. Section 4 describes our analysis of a real data set to further illustrate our method. Section 5 discusses implementation issues concerning posterior inference, MCMC sampling and model selection criteria. Finally Section 6 concludes with a brief discussion.

2. Finite mixture model of linear regressions

Let $(y_i, x_i),\ i = 1, \ldots, n$, be a data set of n observations that come from a heterogeneous population, where y_i is the response variable of the ith observation, and $x_i = (x_{i1}, \ldots, x_{ip})'$ collects the p predictor variables or covariates of the ith observation. We assume that the heterogeneous population consists of M sub-populations or mixture components, and within each sub-population, (y_i, x_i) follows a separate linear regression model. Specifically,

$$y_i|(\rho_m, \beta_m, \sigma_m^2, m = 1, \ldots, M) \sim \sum_{m=1}^{M} \rho_m \cdot N\left(x_i'\beta_m, \sigma_m^2\right), \tag{1}$$

where $\rho = (\rho_1, \ldots, \rho_M)$ is the proportion vector of the M sub-populations, with $\rho_m \geq 0$ and $\sum_{m=1}^{M} \rho_m = 1$. $\beta_m = (\beta_{m1}, \ldots, \beta_{mp})'$ is the coefficient vector for the linear regression in the mth sub-population. σ_m^2 is the corresponding variance of the Gaussian residual errors.

2.1. Two sets of latent variables

As is standard for the mixture model, we introduce a latent variable z_i for each observation i, so that $z_i = m$ indicates that the ith observation comes from the mth sub-population. Thus $P(z_i = m) = \rho_m$, i.e.,

$$z_i \sim \text{Multinomial}(\rho_1, \ldots, \rho_M), \quad \text{and} \quad [y_i|z_i = m] \sim N\left(x_i'\beta_m, \sigma_m^2\right).$$

In modern applications of linear regression models, the number of predictors p can be large, and it is often assumed that the coefficient vector is sparse, i.e., only a small number of its components are different from zero. In other words, only a small number of predictor variables are to be included in the regression model. We assume that this is the case with the