Computational Statistics and Data Analysis 126 (2018) 125–135

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Sparse pathway-based prediction models for high-throughput molecular data

Sangin Lee [a], Youngjo Lee [b], Yudi Pawitan [c,*]

[a] Department of Information and Statistics, Chungnam National University, Republic of Korea
[b] Department of Statistics, Seoul National University, Republic of Korea
[c] Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Sweden

ARTICLE INFO

Article history:
Received 4 October 2017
Received in revised form 26 April 2018
Accepted 26 April 2018
Available online 8 May 2018

Keywords:
Hierarchical likelihood
Overlapping groups
Random-effect model
Structured variable selection

ABSTRACT

Pathway-based prediction problems for high-throughput molecular data motivate the development of sparsity-constrained models with structured predictive variables. Intuitively it is desirable to incorporate the structural information into the model building procedure, potentially for improving both interpretability and prediction performances. Various random-effect models are developed for structured sparse prediction where the predictive variables/genes can be naturally grouped into overlapping groups or pathways. The hierarchical likelihood approach can be used for these random-effect models that impose sparse selection of the overlapping groups as well as further selection within the selected groups. In addition, the approach leads to a unified optimization algorithm for these random-effect models. Extensive numerical studies based on simulated and real breast-cancer data demonstrate that the proposed methods perform well against existing methods that ignore the structural information.

1. Introduction

Sparsity is a crucial element in building reliable prediction models. The main aspect that we shall focus on here is the fact that in the classical approach, such as the penalized regression approach known as the least absolute shrinkage and selection operator (LASSO) procedure (Tibshirani, 1996), sparsity constraint or selection is typically applied to individual features. In other words the selection process ignores any structure or relationship between the features. Intuitively this is not optimal for high-throughput molecular data, because features such as genes are naturally grouped into biological pathways. For example, it may make more sense to decide whether a certain pathway is not involved in a process, so we could drop a whole collection of genes in the pathway. However, we must also account for the well-known pleiotropic effect, where a single gene may belong to multiple biological pathways and its effect may vary across the different pathways. So our goal is to develop a pathway-based feature selection method that allows for overlapping pathways.

There have been efforts to take the prior information about the structure of the predictor variables into account, potentially both for interpretability purpose and prediction accuracy. Yuan and Lin (2006) proposed the group version of LASSO by modifying the penalty function. The smoothly clipped absolute deviation (SCAD) penalty and minimax concave penalty (MCP) have also been extended to the group selection problem (Fan and Li, 2001; Wang et al., 2007; Zhang et al., 2010; Breheny and Huang, 2015). In both of these procedures, the group selection properties are only implicitly determined by the numerical properties of the penalty norms. Therefore it is not obvious how to set-up the penalties so that predictor

[*] Corresponding author.
E-mail address: yudi.pawitan@ki.se (Y. Pawitan).

https://doi.org/10.1016/j.csda.2018.04.012
0167-9473/© 2018 Published by Elsevier B.V.

pathways are allowed to be overlapping. In contrast we shall develop procedures based on random-effect models, where feature selection under the overlapping group structure can be seen in an explicit and transparent way.

To be specific, first consider the following linear regression model where the predictive variables can be divided into K disjoint groups

$$\mathbf{y} = \sum_{k=1}^{K} \mathbf{X}_k \boldsymbol{\gamma}_k + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^T$ is the vector of responses, $\mathbf{X}_k = (\mathbf{x}_{1k}, \ldots, \mathbf{x}_{nk})^T$ is the $n \times p_k$ design matrix for the kth group with the ith-row $\mathbf{x}_{ik}$, $\boldsymbol{\gamma}_k = (\gamma_{k1}, \ldots, \gamma_{kp_k})^T$ is the vector of the corresponding coefficients in the kth group, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ is the vector of random errors. A general group selection can be achieved by minimizing the following criterion:

$$\frac{1}{2n} \sum_{i=1}^{n} \Big(y_i - \sum_{k=1}^{K} \mathbf{x}_{ik}^T \boldsymbol{\gamma}_k \Big)^2 + \sum_{k=1}^{K} p_{\lambda_k}(\|\boldsymbol{\gamma}_k\|_2), \tag{2}$$

where $p_{\lambda_k}(\cdot)$ is a penalty function, λ_k is the regularization parameter for the kth group and $\| \cdot \|_2$ is the ℓ_2-norm operator. Yuan and Lin (2006) proposed the group LASSO by using $p_{\lambda_k}(t) = \lambda_k |t|$ in (2). For nonconvex penalties, the SCAD penalty and MCP were extended to the group selection problem (Breheny and Huang, 2015).

We have previously studied the random-effect model approach for group variable selections and shown that a feature selection that respects the group structure performs well against the standard selection method such as the LASSO (Lee et al., 2015). Compared to the penalized approach, the random-effect model approach is more flexible and transparent; for example, different types of group selection can be obtained by different assumptions on the random effects. In the previous work, the group structure is determined by simple considerations, such as dummy variables associated with one predictor. There is a strong restriction that the groups of predictors cannot overlap.

There have been some works to extend the group LASSO to the overlapping group case. Two versions of group LASSO with overlapping groups – where a variable can belong to multiple groups – have been proposed. In Jenatton et al. (2011) and Yuan et al. (2011), the LASSO penalty is directly applied to the coefficients for each group. The resulting solution consists of an *intersection* of a sub-collection of overlapped groups, i.e., its support is the complement of the union of the estimated null groups. Jacob et al. (2009) introduced the latent-group LASSO approach in which the support of the solution forms a *union* of a sub-collection of predetermined groups. Recently, Zeng and Breheny (2016) extended the latent group LASSO approach to the logistic regression. These methods inherit not only the good properties of group LASSO, but also the disadvantages. For example, as shown in Section 4, they often select more groups than necessary, so tend to produce more false-positive groups.

In this paper, we reformulate the overlapping-group situations by using a generalized linear model with latent coefficients, which covers the two versions above as special cases by using different assumptions for the corresponding coefficients of overlapped variables. We propose the random-effect models that correspond to the above two versions and develop a unified algorithm for them. Moreover, we consider the further selection problems in the selected groups, i.e., *bi-level selection*. As before, the key conceptual advantage of the model-based approach is that it is more transparent and flexible.

The rest of the paper is structured as follows. In Section 2, we introduce the generalized linear model (GLM) with latent coefficients and random-effect models for structured variable selection. In Section 3, we describe the hierarchical likelihood (h-likelihood) approach and the computational algorithm for the proposed random-effect models. Numerical results for simulations and real data analysis are given in Sections 4 and 5. Concluding remarks are provided in Section 6.

2. Random-effect models for pathway-based selection

A key novelty in our approach is the use of random-effect models to impose sparseness and achieve pathway-based selection. To allow general continuous and discrete outcomes, suppose that the response vector $\mathbf{y} = (y_1, \ldots, y_n)^T$ follows a GLM with linear predictors

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} = X_1\beta_1 + \cdots + X_p\beta_p, \tag{3}$$

where $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^T$ is the vector of linear predictors, $\mathbf{X} = (X_1, \ldots, X_p)$ is the $n \times p$ design matrix with the jth-column X_j, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is the vector of regression coefficients. For standard variable selection in GLMs, Lee and Oh (2014) proposed the following random-effect model

$$\beta_j | u_j \sim \mathrm{N}(0, \sigma^2 u_j),\ j = 1, \ldots, p \tag{4}$$

and

$$u_j \sim \mathrm{G}(\alpha),\ j = 1, \ldots, p,$$

where $\mathrm{G}(\alpha)$ is the gamma distribution with mean 1 and variance α. In the model (4), a sparsity can obviously be achieved by $\hat{u}_j = 0$ which leads to $\hat{\beta}_j = 0$, and hence variable selection is achieved in a transparent way. This model is a double