Title
A model-based multithreshold method for subgroup identification

Authors
Wang, Jingli
Li, Jialiang
Li, Yaguang
et al.

Publication Date
2019-02-11

DOI
10.1002/sim.8136

Peer reviewed

Received: 9 March 2018 | Revised: 29 January 2019 | Accepted: 11 February 2019

DOI: 10.1002/sim.8136

RESEARCH ARTICLE

WILEY Statistics in Medicine

A model-based multithreshold method for subgroup identification

Jingli Wang[1] | Jialiang Li[1,2,3] | Yaguang Li[4] | Weng Kee Wong[5]

[1]Department of Statistics and Applied Probability, National University of Singapore, Singapore

[2]Duke University-NUS Graduate Medical School, Singapore

[3]Singapore Eye Research Institute, Singapore

[4]University of Science and Technology of China, Hefei, China

[5]Department of Biostatistics, Fielding School of Public Health, University of California, Los Angeles, Los Angeles, California

Correspondence
Weng Kee Wong, Department of Biostatistics, Fielding School of Public Health, University of California, Los Angeles, Los Angeles, CA 90095-1772.
Email: wkwong@ucla.edu

Funding information
Academic Research Funds, Grant/Award Number: R-155-000-205-114 and R-155-000-195-114; Tier 2 MOE funds in Singapore, Grant/Award Number: MOE2017-T2-2-082, R-155-000-197-112, and R-155-000-197-113; National Institute of General Medical Sciences of the National Institutes of Health, Grant/Award Number: R01GM107639

Thresholding variable plays a crucial role in subgroup identification for personalized medicine. Most existing partitioning methods split the sample based on one predictor variable. In this paper, we consider setting the splitting rule from a combination of multivariate predictors, such as the latent factors, principle components, and weighted sum of predictors. Such a subgrouping method may lead to more meaningful partitioning of the population than using a single variable. In addition, our method is based on a change point regression model and thus yields straight forward model-based prediction results. After choosing a particular thresholding variable form, we apply a two-stage multiple change point detection method to determine the subgroups and estimate the regression parameters. We show that our approach can produce two or more subgroups from the multiple change points and identify the true grouping with high probability. In addition, our estimation results enjoy oracle properties. We design a simulation study to compare performances of our proposed and existing methods and apply them to analyze data sets from a Scleroderma trial and a breast cancer study.

KEYWORDS

change point, factor analysis, PCA, personalized medicine, Scleroderma, subgroup identification

1 | INTRODUCTION

A main aim of precision medicine is to find a treatment that maximizes individual health outcomes. There is a lot of interest in this exciting approach in medicine because of its promise and potential impact in practice. Its rapid rise in interest can be attributed to increasing recognition that (i) the "one size fits all" strategy does not work for many serious diseases, such as cancer, and targeted therapies based on individual traits tend to work better; (ii) it is much more difficult to find a treatment that works for all patients; (iii) risk factors for a disease are likely going to vary among different patients groups; and (iv) recent advances in genomics, computational biology, medical imagining, and regenerative medicine have made targeted therapies more feasible. The overarching goal in precision medicine then is to find subgroups of patients

that respond differentially to different treatment regimens and model the relationships between the response variable and predictors differently across the subgroups.

There are two different goals for subgroup identification in precision medicine, namely, *prognostic* and *predictive* signature developments. Clark et al[1] summarize the two types as follows: a *prognostic* signature is a measurement associated with clinical outcome without therapy or with the application of a standard therapy that patients are likely to receive and it can be thought of as a measure of the natural history of the disease, and a *predictive* signature is a measurement that is associated with response to a particular therapy and that is best evaluated in a randomized clinical trial with a control group. Our focus is on prognostic signature even though it is quite possible that methods developed here may also apply to the predictive setting.

There are many statistical methods for subgroup identification in the literature and some of the most popular ones are based on tree-like partitioning algorithms. Early work includes automatic interaction detection (AID)[2] and theta automatic interaction detection (THAID).[3] The regression tree (CART)[4] algorithm was particularly successful and tree-based methods become more widely used for subgroup identification. A tree recursively partitions the subjects into binary subgroups until certain stopping criterion is met. There are two approaches to find an optimal tree: *prepruning* and *postpruning*. Prepruning relies on some internal stopping criterion to control the size of the tree, such as in the work of Zeileis et al,[5] and postpruning prunes a very large tree to a smaller sized-tree based on some optimality criteria.[6] A classification tree is a tree method where the response takes on discrete or unordered values and a regression tree is a tree method where the response variable is continuous or has ordered discrete values.[7] There are many tree-based classification and regression subgroup identification methods, and they include generalized unbiased interaction detection and estimation (GUIDE),[8,9] model-based recursive partitioning (MOB),[5] interaction tree (IT),[10] subgroup identification based on difference effect search (SIDES),[11] and virtual twins (VT),[12] among others. GUIDE and MOB can be applied to both prognostic and predictive signature developments, but IT, SIDES, and VT are only for predictive signature development. Some researchers have also proposed subgroup identification methods for individualized treatment rules.[13,14] Doove et al[15] and Lipkovich et al[16] provide reviews and compare a few tree-based methods.

Additionally, change point detection (CPD) methods are available in the economic studies[17,18] and may prove useful as an alternative strategy for clustering subjects. Finding a structural break for a covariate in the observed subjects effectively leads to subgrouping. To implement the change point detection is not easy as there are usually two challenges: (i) one needs to decide the number of change points and (ii) one must estimate the change point locations accurately. Earlier authors proposed iterative cumulative-sum methods, which could be computationally intensive. Recent authors adopted the penalization method, which accelerates the change point detection. In this paper, we consider a recently proposed method called two-stage multiple change-point detection (TSMCD) method,[19] which enjoys solid theoretical advantages and nice practical performance. There are other methods for subgroup identification,[20,21] including Bayesian approaches.[22,23]

In general, to apply a CPD approach, such as TSMCD, one first decides on the choice of the thresholding variable for which the change point is sought. Typically, simple thresholding variable is used. For example, in econometric time series modeling, the thresholding variable is simply the time. Existing CPD methods usually take one of the covariates as the thresholding variable, which may be inadequate for partitioning and prediction purposes. The same applies when we choose the splitting variable for a tree method.

This paper proposes a general framework to select a combination of predictor variables as the thresholding variable. We identify a linear combination of covariates to divide the sample into multiple groups according to the change points. This is tantamount to forming parallel change planes in the linear space spanned by the covariates and subjects are thus grouped by the change planes according to their covariate characteristics. We propose a few practical ways to construct the linear combination, including strategies from principal components and factor analysis. The latter was originally developed for the analysis of scores on the mental tests[24] and is now widely applied in many research fields. The factor analysis model can be used for finance,[25] genomics,[26] and neuroscience[27] among many other research fields. The expectation maximization algorithm can be applied to estimate the factor loadings and factors.[28,29] Recently, Fan et al[30] developed a factor adjusted robust multiple testing method for high-dimensional data.

Several of our proposed methods to derive latent variables useful for subgroup identification are based on latent factors and principal components. This approach is motivated from the high-dimensional principal components analysis (PCA) for reducing the data dimension by creating orthogonal eigen-components with ordered importance.[31,32] The PCA can usually help identify meaningful patterns in data and therefore may serve as candidates for the thresholding variable. In what follows, we demonstrate that our methods are practical and, unlike most of the current methods, they can identify a thresholding variable with possibly two or more change points.

After we have specified the thresholding variable, we apply the TSMCD approach to identify the subgroups. We prefer the TSMCD method for two main reasons. First, this method can divide the sample into two or more groups when there is one or more change points. This substantially improves the single change point estimation method, which restricts the separation into only two groups. More flexible partitioning is thus attained when we expect heterogeneous grouping for the study population. Many existing methods, such as AIM-rule and sequential BATTing method,[20] have outstanding performance but they only return two subgroups. Despite their popularity, their grouping may be not as accurate as our method with mixed regression relationship. In addition, these two methods usually do not produce a regression model for subgroups directly. Practitioners usually have to carry out postidentification modeling indirectly. In fact, both AIM-rule and Seq-BATTing methods typically place subjects into signature positive and signature negative subgroups and so only provide information on group membership, but no information on the risk factors and confounder effects for different subgroups. If practitioners need to examine the dependence of the response variable on the predictors for each subgroup, they will need to fit a model within each subgroup based on the splitting results. Second, the TSMCD method has theoretical support and fast computing speed. Under mild technical conditions, the estimated thresholding locations (change points) from TSMCD converge almost surely to the true thresholding locations.[19] Furthermore, the estimated regression coefficients work as well as the oracle estimators, which can only be obtained when the change point structures are known. Establishing similar results for existing recursive partitioning methods or other ad hoc methods has not been fully addressed in the literature yet.

There are several key contributions in our work. First, we propose a new change point detection method for subgroup identification in personalized medicine. Second, we propose a general framework for constructing subgroups via a thresholding variable; our method is both flexible and, new in that, our thresholding variable extends those previously proposed and not limited to a single observed covariate. Our method considers not only the covariates but also a combination of covariates, latent factors, and principal components. Simulation results show that our methods can outperform current methods and provide more meaningful subgroup identification and accurate prediction for all kinds of medical outcomes.

The rest of this paper is organized as follows. In Section 2, we review some existing subgrouping methods, which are to be compared with our proposed methods. In Section 3, we present a piecewise linear regression model with unknown change points and then propose thresholding variable selection methods. Full details of TSMCD will be provided. Section 4 contains simulation studies for assessing the proposed methods. We apply the proposed methods to two medical examples in Section 5. We provide a discussion in Section 6.

2 | A REVIEW OF SELECTED COMPARATOR APPROACHES

A model-based recursive partitioning (MOB)[5] groups the observations into clusters on the basis of covariate values. This method has been appraised by many researchers for its outstanding performance. Specifically, one first fits a model with the entire sample and then performs an M-fluctuation test for the parameter instability with respect to all candidates of thresholding variables. If overall parameter instability is achieved, the variable with the highest parameter instability is issued as a splitting variable. The procedure is repeated in each of the children nodes learned from the preceding step, until convergence. Finally, MOB constructs a tree in which every leaf is associated with a well-fitted submodel. This method is flexible and can deal with continuous or categorical variables, but the measurements of parameter instability for different type of variables have different forms. The performance of MOB may be influenced by dimension of the data because the M-fluctuation tests performed repeatedly may be rather computationally intensive especially when the dimension of the covariates space is high. Another issue is that, sometimes, the splitting variable may be irrelevant practically and hard to interpret. The MOB can be implemented by function `mob` in the `R` package `party`. The argument `method` allows users to select between linear or generalized linear models.

Huang et al[20] proposed two methods for subgroup identification. The first approach is sequential BATTing (seq-BATTing), a multivariate extension of the bootstrapping, and aggregating of thresholds from trees (BATTing), whereas the second approach is AIM-rule, which is a multiplicative rules-based modification of the adaptive index model (AIM).[33] The working models of the mean response for these methods are $m(\boldsymbol{X}) = \beta_0 + \beta_1 w(\boldsymbol{X})$ for prognostic signatures and $m(\boldsymbol{X}) = \beta_0 + \beta_1 w(\boldsymbol{X})u + \beta_2 u$ for predictive signatures, where $m(\boldsymbol{X})$ is the mean response, β_i and $i = 0, 1, 2$ are regression parameters, $w(\boldsymbol{X})$ is the multiplicative signature rule, and u is a treatment variable. Both AIM-rule and Seq-BATTing depend on the original BATTing approach. In the BATTing procedure, there are B single-stub thresholds, which are obtained by a single split on the predictor for B bootstrap data sets, and the best separations are obtained by maximizing the score test statistics. The estimated threshold is a robust estimator (eg, median) for characterizing the

distribution of B single-stub thresholds. Sequential BATTing extends the BATTing procedure with a stepwise method. First, find the thresholding positions for candidate predictors by BATTing and then select the predictor, which can maximize the score test statistics and update the multiplicative signature rule. Then, repeat this procedure on predictors without previously selected predictors until the likelihood ratio test statistics of two adjacent multiplicative signature rules is not significant. For the AIM algorithm, the thresholding variables and positions are obtained by maximizing the score test and then updating the additive signature rule by adding an indicator function of subgroup regions. Since the working models are quite general, both Seq-BATTing and AIM-rule allow continuous, discrete, and censored survival outcomes. However, the working models all assume that the difference of mean responses between two identified subgroups must be a constant β_1. When the difference also depends on covariates (as in our proposed model (1) in the next section), these methods may produce misleading solutions. In practice, both AIM-rule and sequential BATTing methods can be implemented by a function `SubgrpID` in `R` package `SubgrpID`. In particular, AIM-rule can be implemented by `SubgrpID` with argument "`method="AIM.Rule"`", whereas sequential BATTing can be implemented by `SubgrpID` with argument "`method= "Seq.BT"`".

The patient rule induction method (PRIM) was first proposed by Chen et al[34] using the bump hunting algorithm developed in the work of Friedman and Fisher[35] and, since then, have been applied to different biomedical applications. For subgroup identification problems, this approach can be applied to both prognostic signature[36,37] and predictive signatures.[34] Similar to AIM-rule and sequential BATTing, PRIM also avoids making assumptions of specific data generating mechanisms. Moreover, PRIM directly targets subgroups regions rather than indirectly through the estimation of a regression function. The original PRIM method needs a prespecified threshold value, but some upgraded versions only rely on test p-values or other statistics to split the group.[34] When there exist two or more groups, PRIM might not be able to resolve two distinct modes for the response distribution and must be remedied by additional procedures.[38] Compared with MOB, we note that none of AIM-rule, seq-BATTing, and PRIM can produce a fitted model after the splitting step, and users should refit model for identified subgroups to understand the relationship between the response and the predictors. In addition, PRIM can be applied by `SubgrpID` with argument `'method="PRIM"'` in `R` package `SubgrpID`.

Following the terminology in the work of Lipkovich et al,[16] the aforementioned methods may be classified into the following two categories. (i) Global outcome modeling, or scoring-based methods, is to build a composite model to generate a single score (or probability) for each individual, and then use this composite score for further subgroup identification. CART, GUIDE, VT, and a few other learning approaches fall into this category. Specifically, for classification tree methods with a binary outcome, a follow-up cutoff needs to be derived on the probability of the membership at each node to return the final subgroups. (ii) Local modeling, or rule-based modeling, aims for direct subgroup identification of regions in the covariate space without any individual outcome prediction. SIDES, Rule-Fit, PRIM, and seq-BATTing belong to this category. All local modeling methods need to have a further step of regression modeling to produce the individual prediction. This additional procedure, however, may deviate from the purpose of local modeling.

All the aforementioned approaches have found successes in many clinical applications and demonstrated their usefulness for personalized medicine. However, their choice of splitting variable is rather restrictive and is only selected from the set of available covariates. In addition, only MOB allows a statistical model, whereas other three approaches do not directly output a fitted model. We will consider a new model with change point structure and also allow flexible thresholding methods. The model can directly provide a sensible characterization of the underlying data generating mechanism and thus facilitate statistical prediction.

3 | MODEL AND METHODS

Throughout, we let n be the sample size and let y_i be the real-valued univariate response from the ith subject with covariate $\tilde{\boldsymbol{x}}_i = (x_{i1}, \ldots, x_{ip})' \in \mathbb{R}^p, i = 1, \ldots, n$. Without loss of generality, we denote $\boldsymbol{x}_i = (1, \tilde{\boldsymbol{x}}_i')'$ to incorporate the model with an intercept term. We focus on a piecewise linear regression model with s thresholds given by

$$\begin{aligned} y_i &= \sum_{j=1}^{s+1} \boldsymbol{x}_i' \boldsymbol{\beta}_j 1_{\{a_{j-1} < Z_i \le a_j\}} + \epsilon_i, \\ &= \boldsymbol{x}_i' \left[\boldsymbol{\beta}_1 + \sum_{j=1}^{s} (\boldsymbol{\beta}_{j+1} - \boldsymbol{\beta}_j) 1_{\{a_j < Z_i \le a_{s+1}\}} \right] + \epsilon_i, \quad i = 1, \ldots, n, \end{aligned} \tag{1}$$

where $\boldsymbol{\beta}_j \in \mathbb{R}^p, j = 1, \ldots, s+1$, are unknown $(p+1)$-dimensional regression coefficients for $s+1$ subgroups, $s \geq 0$ is the unknown number of thresholds, $a_1, \ldots, a_s$ are unknown structural break locations, and ϵ_i are independent random errors with mean zero and constant variance. The thresholding variable for subject i is Z_i and Section 3.1 provides details on how to determine this variable. To fix idea, we assume response variable is continuous but extensions to other types of responses are possible using a similar construction. For simplification, we denote $\boldsymbol{\theta}_1 = \boldsymbol{\beta}_1, \boldsymbol{\theta}_{j+1} = \boldsymbol{\beta}_{j+1} - \boldsymbol{\beta}_j, j = 1, \ldots, s$, and write the model as

$$\begin{aligned} y_i &= \sum_{j=1}^{s+1} \boldsymbol{x}_i' \boldsymbol{\theta}_j 1_{\{a_{j-1} < Z_i \leq a_{s+1}\}} + \epsilon_i, \\ &= \begin{cases} \boldsymbol{x}_i' \boldsymbol{\theta}_1 + \epsilon_i, & \text{if } a_0 < Z_i \leq a_1, \\ \boldsymbol{x}_i' (\boldsymbol{\theta}_1 + \boldsymbol{\theta}_2) + \epsilon_i, & \text{if } a_1 < Z_i \leq a_2, \\ \cdots & \cdots \\ \boldsymbol{x}_i' \left(\sum_{j=1}^{s+1} \boldsymbol{\theta}_j \right) + \epsilon_i, & \text{if } a_s < Z_i \leq a_{s+1}. \end{cases} \end{aligned} \tag{2}$$

When the number of change points and the their locations are known in model (2), we only need to estimate $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \ldots, \boldsymbol{\theta}_{s+1}')'$ and we can do so using least squares or other familiar regression techniques.

3.1 | Thresholding variable selection

When the thresholding variable Z is given, we may apply the TSMCD method in the work of Li and Jin[19] to find change points and estimate the regression parameters. In practice, the choice of thresholding variable is crucial on the identification and interpretation of the subgroups. We develop four methods to specify the thresholding variable Z in this section. In the first method, we consider individual covariates as the thresholding variable. The second method uses a linear combination of covariates to split the covariate space. The third method is based on multivariate factor analysis. The fourth method is to use the principal components of the covariates as candidates of the thresholding variable.

Method 1. **Single covariate.** Most existing splitting methods such as MOB adopt this traditional approach. Specifically, we can take all available covariates as candidates of thresholding variable Z, ie,

$$Z \in \{X_j, j = 1, \ldots, p\}. \tag{3}$$

This method is intuitive but may be too simple to accommodate complicated grouping mechanism generated according to multiple variables. Combined with TSMCD, we refer this method as 1-TSMCD in the rest of this paper.

Method 2. **Weighted combination.** A thresholding variable may be a linear combination of the covariates. In this case, the thresholding variable can be written as

$$Z = \sum_{j=1}^{p} w_j X_j, \tag{4}$$

where w_j is the weight for variable X_j, which can be specified by users or estimated by some methods. For example, we can take equal weights on the variables if they are equally important for thresholding. Consequently, we can take the average of the covariates as the thresholding variable Z and also implement this in Section 5. This method is denoted as A-TSMCD in the following.

Method 3. Factor analysis (FA). Multivariate data often exhibit similar patterns suggesting the existence of common structure hidden in the observations. Factor analysis is based on a multivariate model in which the observed random variables can be expressed as a sum of a linear combination of certain unobserved *factors* and an error term.[39,40] Suppose $\tilde{\boldsymbol{X}} = (X_1, \ldots, X_p)'$ is a p-dimensional vector with mean μ and covariance matrix $\boldsymbol{\Sigma}$, which is assumed positive definite. Under a factor model, $\tilde{\boldsymbol{X}}$ can be written as

$$\tilde{\boldsymbol{X}} = \boldsymbol{\mu} + \boldsymbol{\Lambda} \boldsymbol{f} + \boldsymbol{\epsilon}, \tag{5}$$

where $\boldsymbol{f} = (f_1, \ldots, f_m)'$ in the vector of unobserved factors ($m < p$), μ is a common variable shared with all components of $\tilde{\boldsymbol{X}}$, $\boldsymbol{\epsilon}$ is the error term, and $\boldsymbol{\Lambda} = (\rho_{ij})$ is a $p \times m$ unknown loading matrix. We assume $\boldsymbol{\epsilon}$ is distributed independently of $\boldsymbol{f}$ and with mean $\mathbf{0}$ and a diagonal covariance matrix $\boldsymbol{\Sigma}_\epsilon$. Model (5) is similar to a multivariate regression model except that regressor $\boldsymbol{f}$ in this case is not observable. When $\mathrm{E}(\boldsymbol{f}) = \mathbf{0}_m$ and $\mathrm{cov}(\boldsymbol{f}) = \boldsymbol{I}_m$, model (5) is called an orthogonal factor model.

Since mean $\boldsymbol{\mu}$ does not affect the covariance of $\tilde{\boldsymbol{X}}$, when $\mathtt{cov}(\boldsymbol{f}) = \boldsymbol{I}$, we have

$$\boldsymbol{\Sigma} = \mathtt{cov}(\tilde{\boldsymbol{X}}) = \mathtt{cov}(\boldsymbol{\Lambda}\boldsymbol{f} + \boldsymbol{\varepsilon}) = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Sigma}_{\varepsilon}.$$

It follows that $\boldsymbol{\Sigma}$ is positive definite, ie,

$$\mathtt{cov}(\tilde{\boldsymbol{X}}, \boldsymbol{f}) = \mathtt{E}\left[(\tilde{\boldsymbol{X}} - \boldsymbol{\mu})(\boldsymbol{f})'\right] = \boldsymbol{\Lambda},$$

and the (i,j)th entry of $\boldsymbol{\Lambda} = (\rho_{ij})$ is the covariance of X_i and f_j. In practice, we assume the distribution to be multivariate normal and obtain the maximum likelihood estimates $\hat{\boldsymbol{\Lambda}}$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon}}$ from the sample of n data points. The factor is then estimated using weighted least squares

$$\hat{\boldsymbol{f}} = \left(\hat{\boldsymbol{\Lambda}}'\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon}}^{-1}\hat{\boldsymbol{\Lambda}}\right)^{-1}\hat{\boldsymbol{\Lambda}}'\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon}}^{-1}(\tilde{\boldsymbol{X}} - \hat{\boldsymbol{\mu}}). \tag{6}$$

It can be shown that this estimator is the minimum variance unbiased linear estimator of $\boldsymbol{f}$. Another estimator of $\boldsymbol{f}$ is the Thomson estimator[41] given by

$$\hat{\boldsymbol{f}} = \hat{\boldsymbol{\Lambda}}'\left(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon}} + \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}'\right)^{-1}(\tilde{\boldsymbol{X}} - \hat{\boldsymbol{\mu}}). \tag{7}$$

Sometimes an orthogonal rotation may be applied to the loading matrix to obtain maximum variation of the squared loadings.

In this third approach of choosing a thresholding variable, the estimated factors are the candidates

$$\boldsymbol{Z} \in \{\hat{\boldsymbol{f}}_1, \ldots, \hat{\boldsymbol{f}}_m\}, \tag{8}$$

where $\hat{\boldsymbol{f}}_j = (\hat{f}_{1j}, \ldots, \hat{f}_{nj})', j = 1, \ldots, m$.

In practice, we implement the factor analysis using the function `factanal` in `R`. The `score` option allows the user to select either the least squares estimator (6) or the Thomson's regression estimator (7). We call these two methods F1-TSMCD and F2-TSMCD, respectively.

Method 4. **Principal component analysis (PCA).** Principal components are widely used in many areas of statistical research such as dimensional reduction. Since the first principal component is the linear combination of covariates with the maximum variance among all possible linear combinations,[40,42] it may be useful to consider the first principal component as the thresholding variable.

Suppose that covariance $\boldsymbol{\Sigma}$ is a positive semidefinite matrix. The first principal component of $\tilde{\boldsymbol{X}}$, denoted as g_1, can be written as the linear combination $g_1 = \boldsymbol{e}_1'\tilde{\boldsymbol{X}}$, with $\boldsymbol{e}_1 \in \mathbb{R}^p$ and $\boldsymbol{e}_1'\boldsymbol{e}_1 = 1$ such that

$$\mathtt{var}\left(\boldsymbol{e}_1'\tilde{\boldsymbol{X}}\right) = \max_{\boldsymbol{e}} \mathtt{var}(\boldsymbol{e}'\tilde{\boldsymbol{X}}) = \max_{\boldsymbol{e}} \boldsymbol{e}'\boldsymbol{\Sigma}\boldsymbol{e}.$$

In practice, we may obtain the PCA solution from a matrix singular value decomposition (SVD)[24,43] using the function `svd` in `R` and obtain the thresholding variable

$$\boldsymbol{Z} = (Z_1, \ldots, Z_n)' = \hat{\boldsymbol{g}}_1 = \tilde{\boldsymbol{x}}\hat{\boldsymbol{e}}_1. \tag{9}$$

The function `prcomp` in `R` performs the PCA on the given data matrix. However, in very high dimensional setting when $p > n$, the classical PCA becomes inconsistent[44-46] and we have to use `svd` instead.

We note that the splitting variable Z can be defined in a way that incorporates the response variable y and covariates $\tilde{\boldsymbol{X}}$. Specifically, some supervised dimension reduction methods can be applied to find the candidates of thresholding variable, such as supervised PCA.[47,48] Operationally, we can also easily add the observed response y as a "covariate" into the empirical covariate matrix and then perform the factor analysis or PCA. This method may be suitable for subgroup identification methods not based on a regression model. However, such analysis cannot be easily included in this paper because the resulting model may invoke the reverse causality issue and hard to interpret. More theoretical study is helpful to explore this idea.

3.2 | Generic TSMCD method

After a decision is made on the thresholding variable Z, we follow up with a TSMCD procedure to estimate the regression coefficients and the number and locations of the thresholds.[19] In the first splitting stage, we set $r = \lfloor k\sqrt{n} \rfloor$ and set $q_n = \lfloor n/r \rfloor - 1$, where k is constant. The value of r will be discussed more in the end of this method. The data sequence is split into $q_n + 1$ segments. The first segment $\mathcal{I}_1 = \{i : Z_i \leq Z_{(n-q_n r)}\}$ involves $n - q_n r$ observations, and the other

segments $\mathcal{I}_j = \{i : Z_{(n-(q_n-j+2)r)} < Z_i \le Z_{(n-(q_n-j+1)r)}\}, j = 2, \dots, q_n + 1$ involves r observations, where $Z_{(1)} \le \cdots \le Z_{(n)}$ are ordered thresholding variables in the sample. Then, the estimator of parameters is given by minimizing

$$\sum_{j=1}^{q_n+1} \frac{b_j}{2n} \sum_{i \in \mathcal{I}_j} \left(y_i - \sum_{k=1}^{j} \boldsymbol{x}_i' \boldsymbol{\theta}_k \right)^2 + \sum_{k=2}^{q_n+1} \Gamma_{\lambda_n, \gamma_n}(||\boldsymbol{\theta}_k||), \tag{10}$$

where b_j is the cluster size of $\mathcal{I}_j$, and $\lambda_n > 0$ and $\gamma_n > 1$ are tuning parameters, $\Gamma_{\lambda_n,\gamma_n}(\|\boldsymbol{\theta}_k\|)$ is a penalty function, and $\|\cdot\|$ is the L_2 norm. The smoothly clipped absolute deviation (SCAD) penalty[49] and the minimax concave penalty[50] are usually recommended to be the penalty function Γ. In this paper, we use the SCAD penalty. Similar to the works of Li and Jin[19] and Jin et al,[51] we set the penalty parameter $\gamma_n = 2.4$ for SCAD penalty. The regularization parameter λ_n can be chosen by minimizing the Bayesian information criterion (BIC)

$$BIC = n \log(RSS/n) + \log(n) DF, \tag{11}$$

where DF is the number of $\theta_{ij} \neq 0, i = 1, \dots, p+1, j = 1, \dots, q_n + 1$, and $RSS = \|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\theta}}\|$, and $\hat{\boldsymbol{\theta}}$ is the estimate of $\boldsymbol{\theta}$ when λ is used in the SCAD penalty.

After we obtain the estimated parameters $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_1', \dots, \hat{\boldsymbol{\theta}}_{q_n+1}')'$, we may then determine the thresholds by checking the nonzero jumps in the estimated coefficients. Let $\hat{\mathcal{A}} = \{j : \hat{\boldsymbol{\theta}}_j \neq \boldsymbol{0}, j = 1, \dots, q_n + 1\}$ and let $\hat{\mathcal{A}}^* = \{j : j \in \hat{\mathcal{A}}, j - 1 \notin \hat{\mathcal{A}}, j = 2, \dots, q_n + 1\} = \{\hat{k}_1, \dots, \hat{k}_{\hat{s}}\}$. If $\hat{s} = 0$, there is no threshold. If $\hat{s} > 0$, the true threshold a_j is highly likely located in $(Z_{(n-(q_n-\hat{k}_j+3)r)}, Z_{(n-(q_n-\hat{k}_j+1)r)}]$, $j = 1, \dots, \hat{s}$. That gives the first-stage estimation.

In the second stage, we refine the estimation procedure and obtain consistent estimators of the thresholds by minimizing the segmented least square errors. To this end, let $\hat{\mathcal{I}}_j = \{i; Z_{(n-(q_n-\hat{k}_j+3)r)} < Z_i \le Z_{(n-(q_n-\hat{k}_j+1)r)}\}$, let $\hat{\mathcal{I}}_{j,\zeta^-} = \{i; Z_{(n-(q_n-\hat{k}_j+3)r)} < Z_i \le \zeta\}$, let $\hat{\mathcal{I}}_{j,\zeta^+} = \{i; \zeta < Z_i \le Z_{(n-(q_n-\hat{k}_j+1)r)}\}$, and let $\hat{b}_{j,\zeta^-}$, $\hat{b}_{j,\zeta^+}$, and $\hat{b}_j$ be the size of the set $\hat{\mathcal{I}}_{j,\zeta^-}$, $\hat{\mathcal{I}}_{j,\zeta^+}$, and $\hat{\mathcal{I}}_j$, respectively,

$$Q_j(\zeta^-, \boldsymbol{\theta}) = \frac{\hat{b}_{j,\zeta^-}}{\hat{b}_j} \sum_{i \in \hat{\mathcal{I}}_{j,\zeta^-}} \left(y_i - \sum_{k=1}^{j} \boldsymbol{x}_i' \boldsymbol{\theta}_k \right)^2,$$

$$Q_j(\zeta^+, \boldsymbol{\theta}) = \frac{\hat{b}_{j,\zeta^+}}{\hat{b}_j} \sum_{i \in \hat{\mathcal{I}}_{j,\zeta^+}} \left(y_i - \sum_{k=1}^{j} \boldsymbol{x}_i' \boldsymbol{\theta}_k \right)^2.$$

Here, ζ^- and ζ^+ represent the left and right limits for a splitting position ζ. The two Q sums represent the left and right sums of squared errors when ζ is used to divide the region. We then estimate each threshold a_j by

$$\hat{a}_j = \text{argmin}_{\zeta \in (Z_{(n-(q_n-\hat{k}_j+3)r)}, Z_{(n-(q_n-\hat{k}_j+1)r)}]} Q_j(\zeta), \tag{12}$$

where $Q_j(\zeta) = \min_{\boldsymbol{\theta}} Q_j(\zeta^-, \boldsymbol{\theta}) + \min_{\boldsymbol{\theta}} Q_j(\zeta^+, \boldsymbol{\theta})$, $\quad j = 1, \dots, \hat{s}$.

After we obtain $\hat{a}_j, j = 1, \dots, \hat{s}$, by (12), it is sensible to use the weighted least squares to compute the final estimates of the coefficients in model (1). Specifically, let $\hat{\mathcal{I}}_j^* = \{i : \hat{a}_{j-1} < Z_i \le \hat{a}_j\}, j = 1, \dots, \hat{s}+1$, $\hat{a}_0 = -\infty$, $\hat{a}_{\hat{s}+1} = +\infty$, and estimate the regression coefficients $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \dots, \boldsymbol{\theta}_{\hat{s}+1}')' = (\theta_1, \dots, \theta_{(p+1)(\hat{s}+1)})'$ by minimizing the penalized least squares

$$\sum_{j=1}^{\hat{s}+1} \frac{\hat{b}_j^*}{2n} \sum_{i \in \hat{\mathcal{I}}_j^*} \left(y_i - \sum_{k=1}^{j} \boldsymbol{x}_i' \boldsymbol{\theta}_k \right)^2 + \sum_{k=1}^{(p+1)(\hat{s}+1)} \Gamma_{\lambda_n, \gamma_n}(|\theta_k|), \tag{13}$$

where $\hat{b}_j^*$ is the size of set $\hat{\mathcal{I}}_j^*$, and the penalty function $\Gamma_{\lambda_n,\gamma_n}(|\cdot|)$ is the same as in (10), $|\theta_k|$ is the absolute value of θ_k.

The performance of $\hat{\boldsymbol{\theta}}$ depends on the initial segment length r, and so it is important to choose a proper r. We choose the optimal r by applying a modified version of BIC. First, apply the splitting stage to the data sequence L times with the common segment length (excluding the first segment) $r_l, l = 1, \dots, L$. We set $r_l = \lfloor k_l \sqrt{n} \rfloor, l = 1, \dots, L$, where k_l takes values from L grid-points in the interval $[0.1, 2]$. For each r_l applying the two-stage generic method (TSMCD), we obtain a set of estimated thresholds $\hat{\mathcal{M}}_l = \{\hat{a}_{1l}, \dots, \hat{a}_{\hat{s}_l l}\}, l = 1, \dots, L$. Then, we use BIC to choose the best index

$$\hat{l} = \texttt{argmin}_{l=1,\dots,L} \{BIC_{\hat{\mathcal{M}}}\},$$

where

$$BIC_{\hat{\mathcal{M}}_l} = n \log(RSS/n) + (p+1)(\hat{s}_l + 1) \log(n), \tag{14}$$

with $\hat{\mathcal{I}}_{j,l} = \{i : \hat{a}_{j-1,l} < Z_i \le \hat{a}_{j,l}\}$, $\hat{a}_{0,l} = -\infty$, $\hat{a}_{\hat{s}_l+1,l} = +\infty$. Our final estimates are results with index $\hat{l}$.

Remark 1. It is known that the ordinary BIC is a liberal measure when the model spaces are large. There are proposed modifications for the criterion when the number of variables increases with the sample size.[52-54] We use the BIC proposed by Wang et al[52] when the dimensions of the covariates are relatively large and modify Equations (11) and (14) by

$$BIC = n\log(RSS/n) + \log(n)DF \cdot C_n, \tag{15}$$

and C_n is a positive constant, which increases to infinity as n increases. The choice of each C_n can be different for these two BIC measures, and usually it takes values $C\log(DF)$, where C is a positive constant.

Remark 2. An important condition for the aforementioned change point detection algorithm to work well is that the proportion for each subgroup should be positive and bounded away from zero (see condition (A1) in the work of Li and Jin[19]). The practical implication of this condition is that we cannot detect a rare group with very small prevalence in the population. Under the technical conditions in the work of Li and Jin,[19] it can be shown that the estimated number of change points $\hat{s}$ is equal to the true s with probability one using the TSMCD algorithm. In addition, the estimated locations of the change points are also consistent to the true change points when the sample size is large.[19] These asymptotic results provide solid theoretical justification to the subgrouping results.

4 | SIMULATION STUDIES

We now perform simulations to compare the performances of the various proposed algorithms, ie, 1-TSMCD, A-TSMCD, F1-TSMCD, F2-TSMCD, and PC-TSMCD, with existing subgrouping methods including AIM-rule, seq-BATTing, PRIM, and MOB. All our simulation programs are run using the `R` software. Following suggestions in the work of Huang et al,[20] the cross-validation is implemented when using the `SubgrpID` package. If the cross-validation p-value < 0.05, we claim that the subgroup is identified.

Our simulation setup consists of eight cases designed to evaluate the performance of different subgrouping methods. In all cases, we repeat the simulation 500 times using sample sizes $n = 300$ and 500, and all models have a normally distributed error term ϵ_i, with mean 0 and variance 0.25. Unlike previous simulation studies, where the thresholding variable is usually one of the observed covariates, we form subgroups either by splitting the unobserved latent variables (Cases I to IV) or by selecting a fixed linear combination of observed covariates (Cases V to VIII).

For Cases I to IV, we apply the factor analysis method to estimate the thresholding variable and generate the covariates $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})'$ from the following factor model:

$$\boldsymbol{x}_i = \boldsymbol{\Lambda}\boldsymbol{f}_i + \boldsymbol{\epsilon}_i, i = 1, \ldots, n.$$

Here, $\boldsymbol{\Lambda}$ is a $p \times m$ loading matrix where each row is generated from the standard multivariate normal distribution, the factors $\boldsymbol{f}_i = (f_{i1}, \ldots, f_{im})'$ are generated from the multivariate normal distribution $N(\boldsymbol{0}, \boldsymbol{I}_m)$, and the stochastic error $\boldsymbol{\epsilon}_i$ is generated from the multivariate normal distribution $N(\boldsymbol{0}, 0.1\boldsymbol{I}_p)$. There are $p = 10$ covariates and $m = 3$ factors.

Case I: We generate data from the model without any change point (ie, no subgroups)

$$y_i = 1 + x_{i1} + \frac{1}{2}x_{i2} + 2(x_{i3} + x_{i4}) + \epsilon_i,$$

for $i = 1, \ldots, n$.

Case II: We generate data from the model

$$y_i = 1 + x_{i1} + x_{i2}1_{\{f_{i1} \le 0\}} + 2(x_{i3} + x_{i4})1_{\{f_{i1} > 0\}} + \epsilon_i,$$

for $i = 1, \ldots, n$. This model involves only one change point (two subgroups). The threshold 0 is chosen so that we have balanced groups in this example. Figure 1 shows how the mean response depends differently on each of the covariates when a data set of size $n = 500$ is generated from the model.

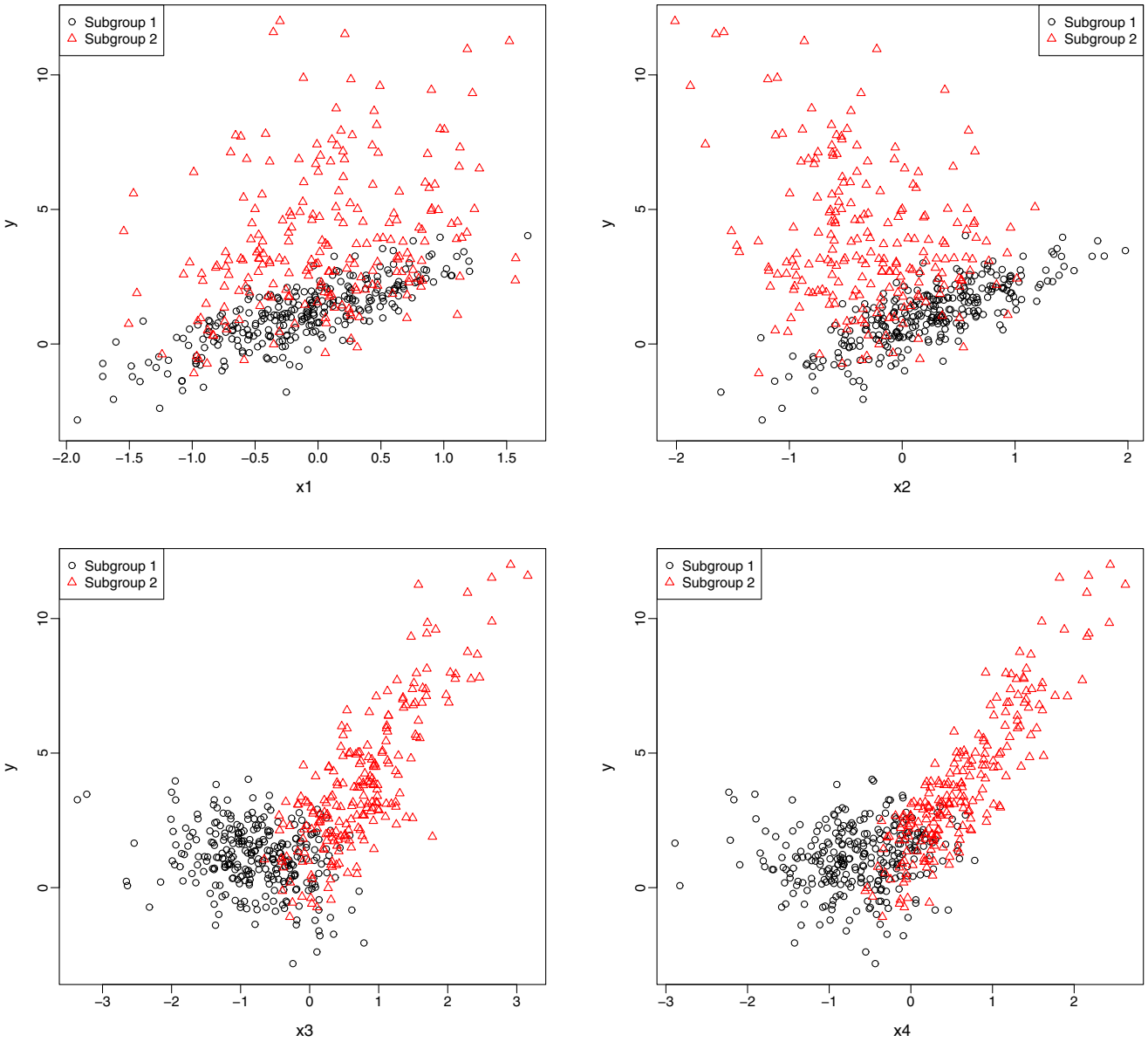

FIGURE 1 Scatter plots of the response variable y and covariates for different subgroups under case II: "Subgroup 1" is under condition $f_{i1} \leq 0$ and "Subgroup 2" is under condition $f_{i1} > 0$ [Colour figure can be viewed at wileyonlinelibrary.com]

Case III: We generate data from the model

$$y_i = 1 + x_{i1} - x_{i2}1_{\{f_{i1} \leq a_1\}} + \left(2x_{i3} + \frac{1}{2}x_{i4}\right)1_{\{a_1 < f_{i1} \leq a_2\}} + 1_{\{f_{i1} > a_2\}} + \epsilon_i,$$

for $i = 1, \ldots, n$, where $a_1 = -0.5$ and $a_2 = 0.5$ are two change points. The two thresholds are chosen so that the three groups are of roughly the same size and the intermediate group is of a relatively greater size. Figure 2 shows how the response variable depends on each of the covariates when a data set of size $n = 500$ is generated from the model.

Case IV: We generate data from the following model that contains a treatment variable

$$y_i = 1 + x_{i1} + (u_i + x_{i2})1_{\{f_{i1} \leq 0\}} + 2(x_{i3} + x_{i4})1_{\{f_{i1} > 0\}} + \epsilon_i,$$

where $u_i \sim Bernoulli(0.5)$ is a binary treatment indicator. This model involves only one change point and hence two subgroups. Figure 3 shows how the response variable depends on each of the covariates when a data set of size $n = 500$ is generated from the model.

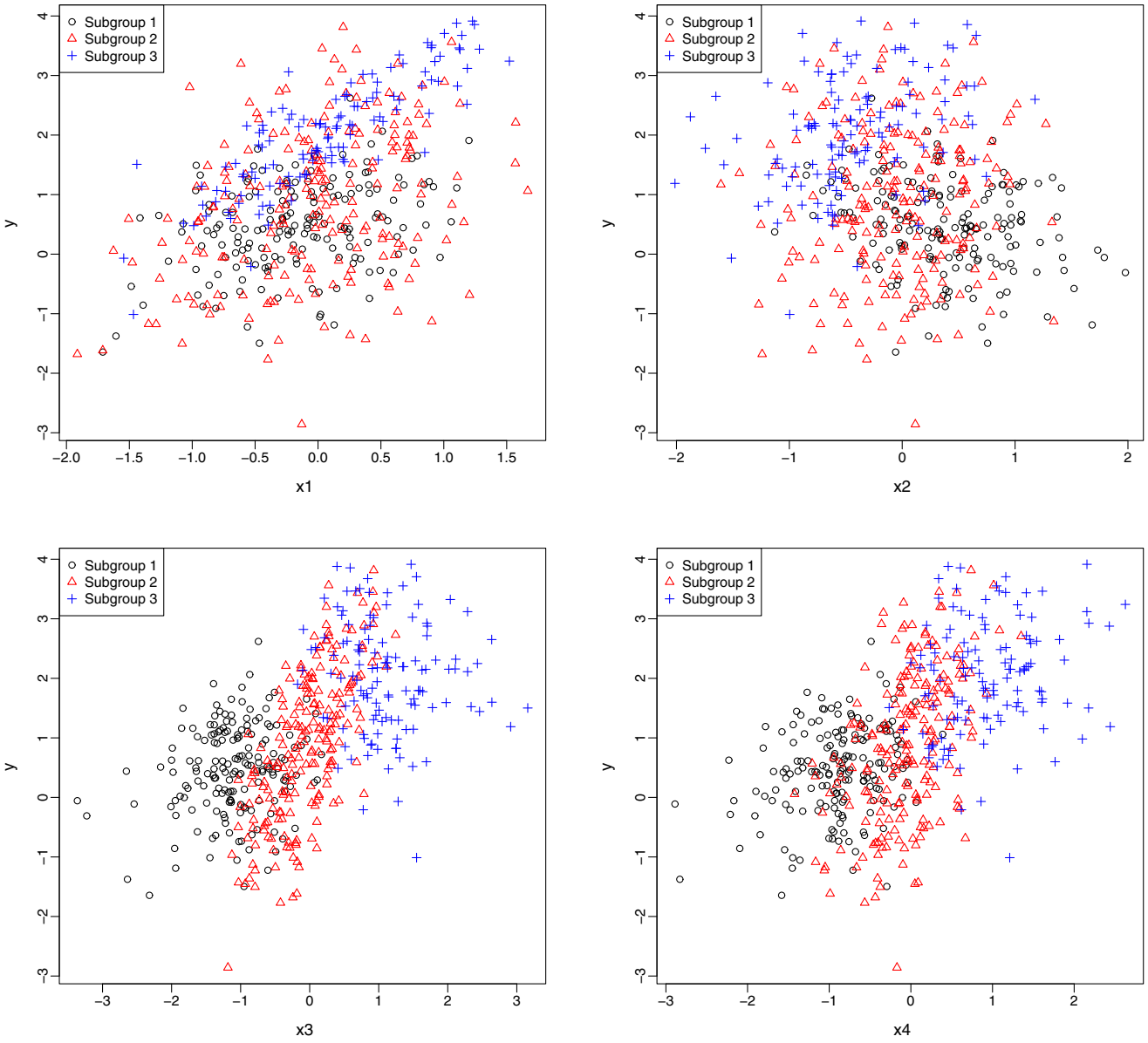

FIGURE 2 Scatter plots of the response variable y and covariates for different subgroups under case III: "Subgroup 1" is under condition $f_{i1} \leq -0.5$, "Subgroup 2" is under condition $-0.5 < f_{i1} \leq 0.5$ and "Subgroup 3" is under condition $f_{i1} > 0.5$ [Colour figure can be viewed at wileyonlinelibrary.com]

Cases V to VIII are designed to examine the performance of PCA-based thresholding. The thresholding variable could be some combination of covariates. Here, we simply use the mean of the covariates as thresholding variable. For Cases V, VI, and VIII, we have $p = 6$ and generate the covariates $x_{i1}, \ldots, x_{i6}, i = 1, \ldots, n$ from a multivariate normal distribution with mean 0 and a covariance matrix $\mathbf{\Sigma}_1$ given by

$$\mathbf{\Sigma}_1 = \begin{bmatrix} 1 & 0.36 & 0.285 & 0.248 & 0.229 & 0.219 \\ 0.36 & 1 & 0.36 & 0.285 & 0.248 & 0.229 \\ 0.285 & 0.36 & 1 & 0.36 & 0.285 & 0.248 \\ 0.249 & 0.285 & 0.36 & 1 & 0.36 & 0.285 \\ 0.229 & 0.248 & 0.285 & 0.36 & 1 & 0.36 \\ 0.219 & 0.229 & 0.248 & 0.285 & 0.36 & 1 \end{bmatrix}.$$

For Case VII, we generate a high-dimensional vector of $p = 50$ covariates $x_{i1}, \ldots, x_{i50}, i = 1, \ldots, n$, from the multivariate normal distribution with zero mean. The covariance of the first six variables is $\mathbf{\Sigma}_1$ and the remaining $p - 6$ variables are uncorrelated with unit variances.

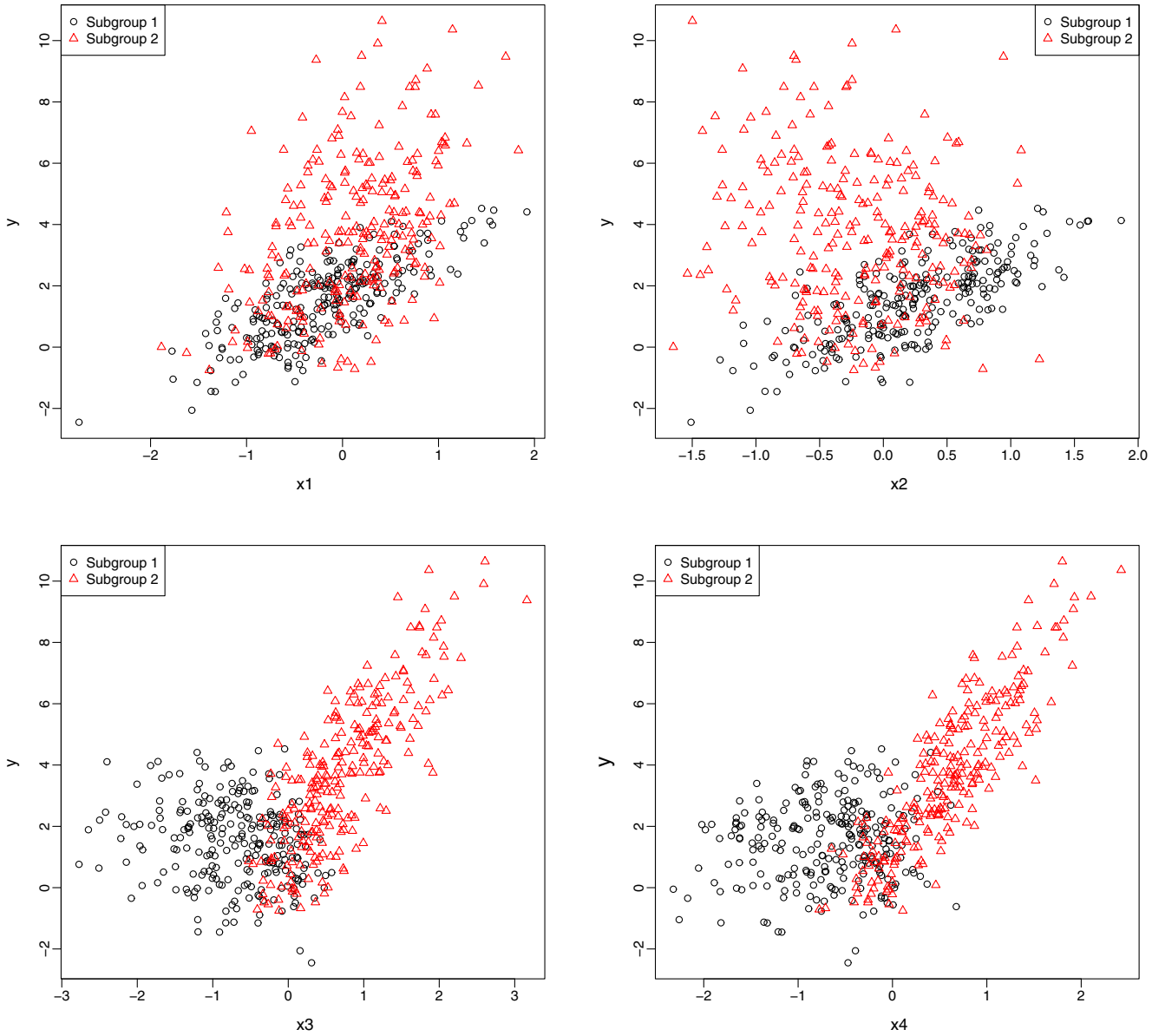

FIGURE 3 Scatter plots of the response variable y and covariates for different subgroups under case IV: "Subgroup 1" is under condition $f_{i1} \leq 0$ and "Subgroup 2" is under condition $f_{i1} > 0$ [Colour figure can be viewed at wileyonlinelibrary.com]

Case V: We generate data from the following model with only one group:

$$y_i = 1 + x_{i1} - 2x_{i2} + x_{i3} + \frac{1}{2}x_{i4} + 0 \cdot x_{i5} + 0 \cdot x_{i6} + \epsilon_i. \tag{16}$$

Case VI: We generate data from the model

$$y_i = -1 + x_{i1} - 2(1 + x_{i2})1_{\{z_i \leq a_1\}} + 2(x_{i2} + x_{i3})1_{\{a_1 < z_i \leq a_2\}} + 3x_{i4}1_{\{z_i > a_2\}} + x_{i5} + \epsilon_i,$$

with $z_i = \sum_{j=1}^{6} x_{ij}, i = 1, \ldots, n$, and $a_1 = -2$ and $a_2 = 2$ are the 30% and 70% quantiles of z_i. Figure 4 shows how the response variable depends on each of the covariates when a data set of size $n = 500$ is generated from the model.

Case VII: We generate data from the model

$$y_i = 1 + x_{i1} + x_{i2}1_{\{z_i \leq 0\}} + 2(x_{i3} + x_{i4})1_{\{z_i > 0\}} + \epsilon_i,$$

with $z_i = \sum_{j=1}^{6} x_{ij}, i = 1, \ldots, n$. Figure 5 shows how the response variable depends on each of the covariates when a data set of size $n = 500$ is generated from the model.

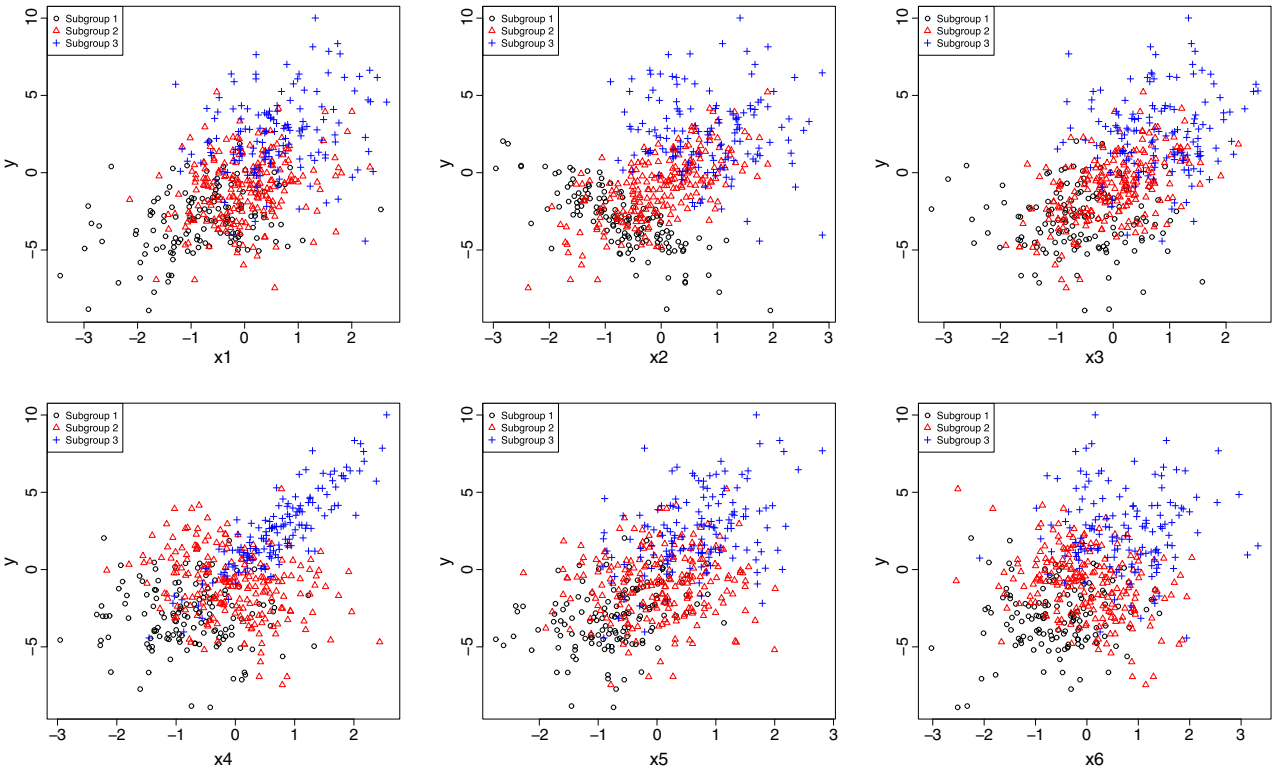

FIGURE 4 Scatter plots of the response variable y and covariates for different subgroups under case VI: "Subgroup 1" is under condition $\sum_{j=1}^{6} x_{ij} \leq -2$, "Subgroup 2" is under condition $-2 < \sum_{j=1}^{6} x_{ij} \leq 2$ and "Subgroup 3" is under condition $\sum_{j=1}^{6} x_{ij} > 2$ [Colour figure can be viewed at wileyonlinelibrary.com]

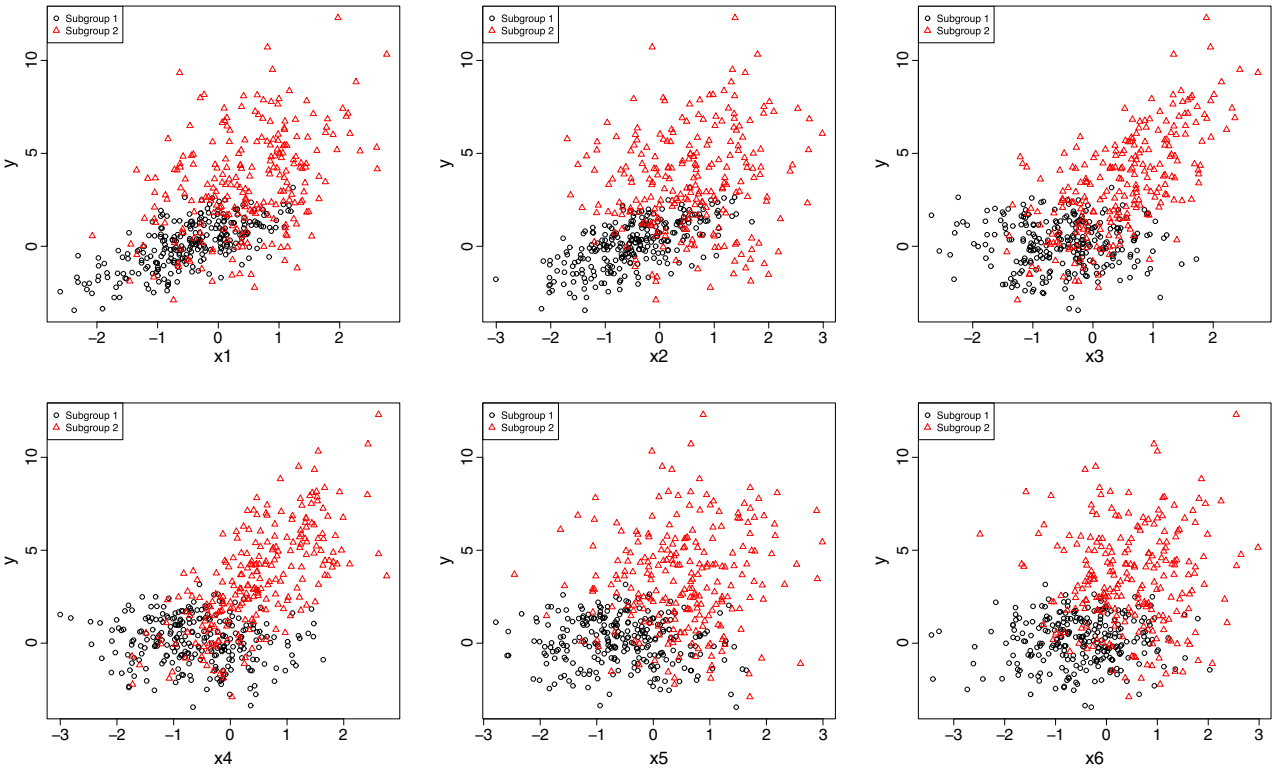

FIGURE 5 Scatter plots of the response variable y and covariates for different subgroups under case VII: "Subgroup 1" is under condition $\sum_{j=1}^{6} x_{ij} \leq 0$ and "Subgroup 2" is under condition $\sum_{j=1}^{6} x_{ij} > 0$ [Colour figure can be viewed at wileyonlinelibrary.com]

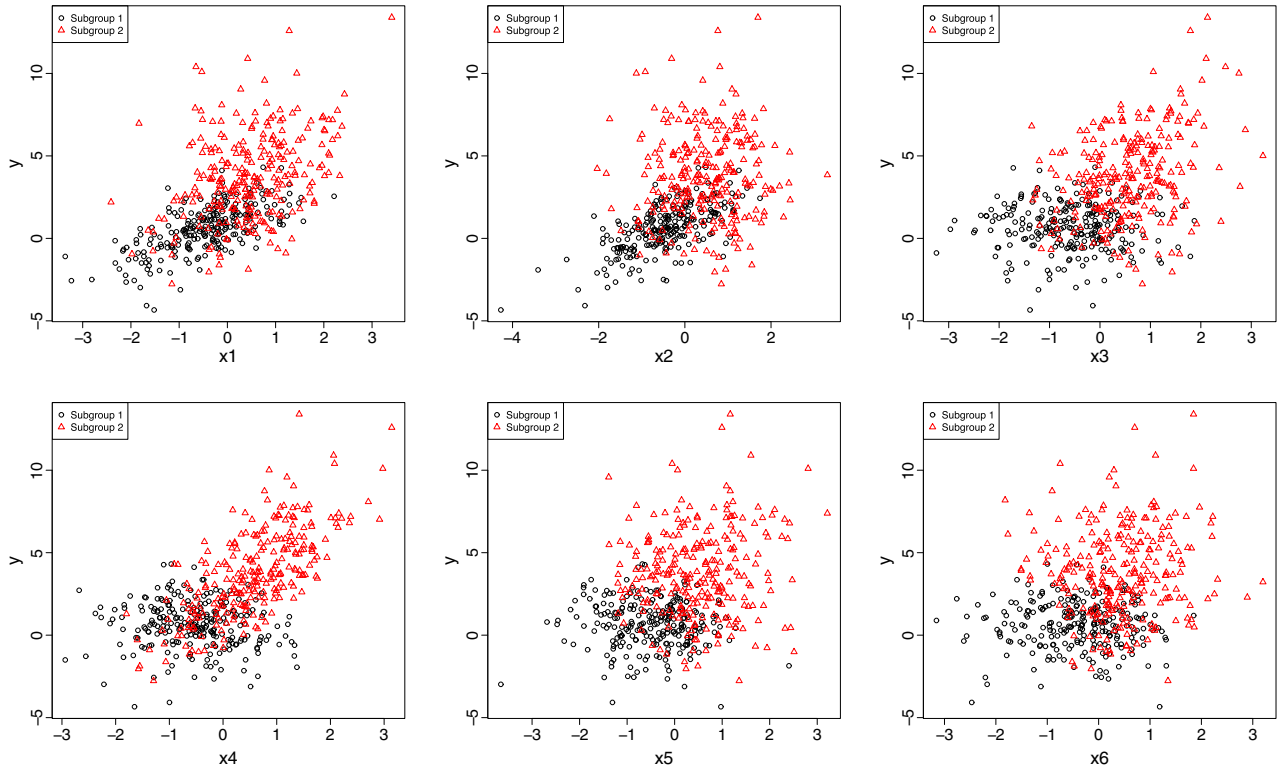

FIGURE 6 Scatter plots of the response variable y and covariates for different subgroups under case VIII: "Subgroup 1" is under condition $\sum_{j=1}^{6} x_{ij} \leq 0$ and "Subgroup 2" is under condition $\sum_{j=1}^{6} x_{ij} > 0$ [Colour figure can be viewed at wileyonlinelibrary.com]

Case VIII: We generate data from the model with a treatment variable

$$y_i = 1 + x_{i1} + x_{i2} 1_{\{z_i \leq 0\}} + 2(x_{i3} u_i + x_{i4}) 1_{\{z_i > 0\}} + u_i + \epsilon_i,$$

where $z_i = \sum_{j=1}^{6} x_{ij}, i = 1, \ldots, n$ and u_i is a treatment variable that follows a Bernoulli distribution with parameter 0.5. Figure 6 shows how the response variable depends on each of the covariates when a data set of size $n = 500$ is generated from the model.

Table 1 shows the summarized simulation results for cases I to VIII, which also include results from the oracle TSMCD method with known true thresholding variable (but with unknown number and location of change points). We note that AIM-rule, seq-BATTing, and PRIM can only split the sample into two subgroups, hence, in Table 1, the identified subgroup numbers $\hat{s} + 1$ for these methods are always two. We also calculate the rate of true subgroup numbers identified (RT) and the positive predictive value (PPV) for each subgroup, and we summarized the results in Table 2.

In case I and case V, there is no subgroup and an appropriate subgrouping method should not recommend to divide the subjects into more than one group. All methods have comparable MSE for both training set and test set. From Table 2, we can find that MOB, F1-TSMCD, F2-TSMCD, and PC-TSMCD can identify the group with very high probability when there is only one group. Since AIM-rule, seq-BATTing, and PRIM all tend to split the sample into two subgroups, their RTs are relatively lower.

For case II, the true subgroup number $s + 1$ is 2. Table 1 shows that F1-TSMCD (factor analysis by weighted least square estimator), F2-TSMCD (factor analysis by Thomson's regression estimator) and MOB achieve much smaller MSE for both training and test sample than AIM-rule, seq-BATTing, and PRIM. Furthermore, both F2-TSMCD and F1-TSMCD are slightly better than MOB in terms of prediction error. While our proposed F1-TSMCD and F2-TSMCD can give closer value to the true number of subgroups, MOB gives a slightly larger groups. In case III, we make a similar observation as in case II.

TABLE 1 Summary of results for case I to VIII. MSE.tr(sd): averaged mean squared error for response among 500 simulations(standard deviation) for the training sample; MSE.te(sd): averaged mean squared error for response among 500 simulations(standard deviation) for the test sample; s+1: true number of subgroups; $\hat{s}+1$ (sd): averaged estimated number of subgroups and standard deviation; AIM-rule: multiplicative rules-based modification of the adaptive index model; seq-BT: sequential BATTing; PRIM: patient rule induction method; MOB: model-based recursive partitioning method; F1-TSMCD: TSMCD with factor analysis (weighted least square estimator); F2-TSMCD: TSMCD with factor analysis (regression estimator); PC-TSMCD: TSMCD with principal components; LM: linear regression model; TSMCD: TSMCD with known true thresholding variable

		n = 300			n = 500		
	Method	**MSE.tr(sd)**	**MSE.te(sd)**	$\hat{s}+1$ **(sd)**	**MSE.tr(sd)**	**MSE.te(sd)**	$\hat{s}+1$ **(sd)**
Case I	AIM-rule	0.2228(0.0212)	0.2824(0.0440)	2(0)	0.2346(0.0181)	0.2671(0.0314)	2(0)
s+1=1	seq-BT	0.2231(0.0216)	0.2818(0.0447)	2(0)	0.2392(0.0153)	0.2521(0.0295)	2(0)
	PRIM	0.2228(0.0214)	0.2814(0.0440)	2(0)	0.2344(0.0180)	0.2671(0.0311)	2(0)
	MOB	0.2354(0.0230)	0.2640(0.0403)	1.0180(0.1331)	0.2414(0.0189)	0.2597(0.0306)	1.0520(0.2222)
	F1-TSMCD	0.2422(0.0228)	0.2591(0.0392)	1(0)	0.2461(0.0191)	0.2545(0.0292)	1(0)
	F2-TSMCD	0.2422(0.0228)	0.2591(0.0392)	1(0)	0.2461(0.0191)	0.2545(0.0292)	1(0)
	LM	0.2361(0.0223)	0.2660(0.0405)	1(0)	0.2425(0.0187)	0.2584(0.0297)	1(0)
Case II	AIM-rule	0.7000(0.1394)	0.8751(0.2215)	2(0)	0.7482(0.1188)	0.8551(0.1683)	2(0)
s+1=2	seq-BT	0.6340(0.0975)	0.7770(0.1655)	2(0)	0.6908(0.0800)	0.7830(0.1344)	2(0)
	PRIM	0.4910(0.1625)	0.7273(0.3739)	2(0)	0.5031(0.1424)	0.6344(0.2618)	2(0)
	MOB	0.3283(0.0534)	0.5366(0.8617)	2.0160(0.1256)	0.3406(0.0415)	0.5859(1.0362)	2.0860(0.2806)
	F1-TSMCD	0.2966(0.0424)	0.3672(0.0972)	2(0)	0.2422(0.0228)	0.2591(0.0392)	2.0040(0.0632)
	F2-TSMCD	0.2850(0.0392)	0.3664(0.1101)	2(0)	0.2422(0.0228)	0.2591(0.0392)	2.0020(0.0447)
	TSMCD	0.2394(0.0244)	0.2714(0.0520)	2(0)	0.2441(0.0202)	0.2616(0.0357)	2(0)
Case III	AIM-rule	0.4198(0.0455)	0.5075(0.0746)	2(0)	0.4289(0.0359)	0.4856(0.0620)	2(0)
s+1=3	seq-BT	0.4496(0.0485)	0.5308(0.0783)	2(0)	0.4592(0.0402)	0.5115(0.0616)	2(0)
	PRIM	0.4228(0.0425)	0.5128(0.0732)	2(0)	0.4336(0.0324)	0.4909(0.0549)	2(0)
	MOB	0.3609(0.0536)	0.5949(0.2978)	2.5824(0.5501)	0.3362(0.0347)	0.5957(0.3027)	3.2784(0.5839)
	F1-TSMCD	0.3659(0.0667)	0.4577(0.0913)	2.7118(0.5256)	0.3462(0.0359)	0.4086(0.0639)	2.9780(0.1939)
	F2-TSMCD	0.3406(0.0553)	0.4309(0.0817)	2.7588(0.4373)	0.3278(0.0371)	0.3895(0.0649)	2.9560(0.2494)
	TSMCD	0.2394(0.0244)	0.2858(0.0448)	3(0)	0.2435(0.0193)	0.2673(0.0324)	3.0000(0.0627)
Case IV	AIM-rule	1.4016(0.3054)	1.8241(0.5061)	2(0)	1.4925(0.2596)	1.7527(0.3516)	2(0)
s+1=2	seq-BT	1.2945(0.4429)	1.6795(0.7902)	2(0)	1.3664(0.4480)	1.5716(0.5499)	2(0)
	PRIM	1.1633(0.3905)	1.5037(0.5387)	2(0)	1.2062(0.3863)	1.4032(0.4844)	2(0)
	MOB	0.3584(0.0614)	0.5658(0.7032)	2.0151(0.1220)	0.3633(0.0471)	0.9574(1.8822)	2.1811(0.3904)
	F1-TSMCD	0.3176(0.0479)	0.4200(0.1245)	2.0019(0.0434)	0.3239(0.0368)	0.3853(0.0784)	2.0019(0.0434)
	F2-TSMCD	0.3095(0.0477)	0.4205(0.1343)	2(0)	0.3129(0.0361)	0.3793(0.0835)	2(0)
	TSMCD	0.2375(0.0239)	0.2736(0.0530)	2(0)	0.2418(0.0179)	0.2637(0.0360)	2(0)
Case V	AIM-rule	0.2323(0.0234)	0.2695(0.0387)	2(0)	0.2416(0.0190)	0.2620(0.0283)	2(0)
s+1=1	seq-BT	0.2376(0.0199)	0.2531(0.0379)	2(0)	0.2445(0.0150)	0.2524(0.0276)	2(0)
	PRIM	0.2324(0.0236)	0.2683(0.0382)	2(0)	0.2416(0.0190)	0.2622(0.0278)	2(0)
	MOB	0.2389(0.0251)	0.2621(0.0406)	1.0840(0.2848)	0.2453(0.0202)	0.2566(0.0315)	1.1040(0.3427)
	PC-TSMCD	0.2429(0.0240)	0.2564(0.0361)	1(0)	0.2482(0.0192)	0.2549(0.0270)	1(0)
	LM	0.2409(0.0239)	0.2585(0.0367)	1(0)	0.2469(0.0194)	0.2562(0.0273)	1(0)
Case VI	AIM-rule	3.2233(0.3248)	3.6864(0.5603)	2(0)	3.2863(0.2826)	3.5802(0.4791)	2(0)
s+1=3	seq-BT	3.2220(0.3775)	3.7026(0.5997)	2(0)	3.2895(0.3400)	3.5941(0.5329)	2(0)
	PRIM	3.1247(0.3358)	3.6877(0.5680)	2(0)	3.2676(0.2745)	3.6152(0.4397)	2(0)
	MOB	2.3196(0.3261)	4.7345(2.2454)	3.0040(0.5182)	2.1001(0.2307)	4.1435(2.1897)	4.1235(0.6319)
	PC-TSMCD	0.4215(0.1789)	1.1065(0.6282)	3.0260(0.1593)	0.4426(0.1770)	1.0047(0.4892)	3.0490(0.2161)
	TSMCD	0.3361(0.1966)	0.4414(0.2584)	3.0060(0.0773)	0.3644(0.2112)	0.4333(0.2505)	3.0412(0.2085)
Case VII	AIM-rule	0.7985(0.1466)	2.9443(1.6657)	2(0)	0.9023(0.1287)	2.1052(0.9457)	2(0)
s+1=2	seq-BT	0.7488(0.1160)	2.5086(1.0276)	2(0)	0.8727(0.1076)	1.7931(0.3189)	2(0)
	PRIM	0.6503(0.1302)	2.1329(4.7812)	2(0)	0.7519(0.1210)	1.5892(0.3286)	2(0)
	MOB	0.7110(0.3028)	2.2404(2.1449)	1.8700(0.3366)	0.7261(0.0983)	1.6432(0.2799)	1.9980(0.0447)
	PC-TSMCD	0.6988(0.3268)	1.2524(0.5547)	1.9360(0.2450)	0.5805(0.1215)	0.9698(0.3437)	2(0)
	TSMCD	0.2430(0.0229)	0.2800(0.0740)	2(0)	0.2473(0.0525)	0.2772(0.0850)	2(0)
Case VIII	AIM-rule	1.0687(0.1906)	1.3090(0.3066)	2(0)	1.0121(0.134)	1.1971(0.212)	2(0)
s+1=2	seq-BT	0.9997(0.1879)	1.2058(0.2756)	2(0)	1.0811(0.1321)	1.2691(0.2315)	2(0)
	PRIM	1.1967(0.1995)	1.4754(0.3293)	2(0)	1.2190(0.1747)	1.3924(0.2652)	2(0)
	MOB	0.7774(0.1206)	1.0623(0.2534)	2(0)	0.6805(0.0999)	1.8066(1.5495)	3.1545(0.6212)
	PC-TSMCD	0.4870(0.1988)	0.7233(0.3277)	2.1611(0.3779)	0.2991(0.1632)	0.4518(0.2380)	2.0255(0.1576)
	TSMCD	0.2771(0.1887)	0.3298(0.2403)	2.0074(0.0858)	0.2664(0.1538)	0.2993(0.1793)	2(0)

TABLE 2 Summary of results for case I to VIII. RT: rate of true subgroups identified; PPV-1, PPV-2, PPV-3: positive predictive values for groups 1, 2, and 3; AIM-rule: multiplicative rules-based modification of the adaptive index model; seq-BT: sequential BATTing; PRIM: patient rule induction method; MOB: model-based recursive partitioning method; F1-TSMCD: TSMCD with factor analysis (weighted least square estimator); F2-TSMCD: TSMCD with factor analysis (regression estimator); PC-TSMCD: TSMCD with principal components; TSMCD: TSMCD with known true thresholding variable

		n=300								n=500							
	Method	RT		PPV-1		PPV-2		PPV-3		RT		PPV-1		PPV-2		PPV-3	
	Method	train	test	train	test	train	test	train	test	train	test	train	test	train	test	train	test
Case I	AIM-rule	0.569	0.577	-	-	-	-	-	-	0.560	0.566	-	-	-	-	-	-
s+1=1	seq-BT	0.620	0.630	-	-	-	-	-	-	0.611	0.619	-	-	-	-	-	-
	PRIM	0.565	0.575	-	-	-	-	-	-	0.557	0.568	-	-	-	-	-	-
	MOB	0.996	0.996	-	-	-	-	-	-	0.991	0.991	-	-	-	-	-	-
	F1-TSMCD	1	1	-	-	-	-	-	-	1	1	-	-	-	-	-	-
	F2-TSMCD	1	1	-	-	-	-	-	-	1	1	-	-	-	-	-	-
Case II	AIM-rule	0.728	0.731	0.458	0.465	0.996	0.991	-	-	0.728	0.726	0.461	0.457	0.995	0.992	-	-
s+1=2	seq-BT	0.76	0.76	0.528	0.527	0.99	0.987	-	-	0.752	0.75	0.514	0.509	0.99	0.989	-	-
	PRIM	0.852	0.844	0.77	0.752	0.938	0.934	-	-	0.859	0.853	0.785	0.772	0.936	0.932	-	-
	MOB	0.876	0.868	0.878	0.867	0.875	0.868	-	-	0.862	0.855	0.874	0.862	0.85	0.846	-	-
	F1-TSMCD	0.92	0.899	0.912	0.893	0.928	0.912	-	-	0.921	0.904	0.912	0.898	0.929	0.915	-	-
	F2-TSMCD	0.937	0.908	0.933	0.908	0.94	0.916	-	-	0.936	0.919	0.927	0.912	0.943	0.929	-	-
	TSMCD	0.995	0.986	0.995	0.987	0.995	0.986	-	-	0.997	0.992	0.997	0.994	0.997	0.99	-	-
Case III	AIM-rule	0.447	0.447	0.363	0.369	0.991	0.988	-	-	0.448	0.447	0.368	0.37	0.991	0.99	-	-
s+1=3	seq-BT	0.405	0.401	0.249	0.246	0.996	0.995	-	-	0.403	0.4	0.247	0.244	0.995	0.996	-	-
	PRIM	0.477	0.473	0.441	0.436	0.992	0.991	-	-	0.47	0.467	0.422	0.417	0.994	0.994	-	-
	MOB	0.572	0.571	0.914	0.91	0.461	0.463	0.362	0.363	0.579	0.579	0.88	0.88	0.409	0.412	0.49	0.489
	F1-TSMCD	0.751	0.733	0.947	0.931	0.667	0.647	0.646	0.636	0.839	0.82	0.933	0.919	0.76	0.739	0.842	0.826
	F2-TSMCD	0.78	0.754	0.935	0.916	0.722	0.694	0.691	0.676	0.855	0.837	0.914	0.902	0.803	0.784	0.86	0.843
	TSMCD	0.989	0.98	0.997	0.988	0.984	0.977	0.986	0.975	0.991	0.986	0.998	0.993	0.987	0.985	0.987	0.98
Case IV	AIM-rule	0.585	0.575	0.643	0.64	0.519	0.512	-	-	0.577	0.566	0.717	0.709	0.431	0.422	-	-
s+1=2	seq-BT	0.648	0.648	0.58	0.58	0.718	0.713	-	-	0.646	0.641	0.555	0.553	0.736	0.733	-	-
	PRIM	0.655	0.65	0.677	0.675	0.633	0.626	-	-	0.657	0.656	0.672	0.67	0.643	0.642	-	-
	MOB	0.876	0.865	0.901	0.889	0.851	0.842	-	-	0.847	0.84	0.888	0.876	0.806	0.804	-	-
	F1-TSMCD	0.922	0.901	0.919	0.904	0.925	0.904	-	-	0.925	0.909	0.926	0.914	0.923	0.907	-	-
	F2-TSMCD	0.933	0.908	0.933	0.914	0.932	0.907	-	-	0.937	0.919	0.939	0.925	0.934	0.916	-	-
	TSMCD	0.996	0.988	0.997	0.991	0.995	0.985	-	-	0.998	0.993	0.998	0.993	0.998	0.992	-	-

(Continues)

TABLE 2 (Continued) Summary of results for case I to VIII. RT: rate of true subgroups identified; PPV-1, PPV-2, PPV-3: positive predictive values for groups 1, 2, and 3; AIM-rule: multiplicative rules-based modification of the adaptive index model; seq-BT: sequential BATTing; PRIM: patient rule induction method; MOB: model-based recursive partitioning method; F1-TSMCD: TSMCD with factor analysis (weighted least square estimator); F2-TSMCD: TSMCD with factor analysis (regression estimator); PC-TSMCD: TSMCD with principal components; TSMCD: TSMCD with known true thresholding variable

		n=300								n=500							
	Method	**RT**		**PPV-1**		**PPV-2**		**PPV-3**		**RT**		**PPV-1**		**PPV-2**		**PPV-3**	
	Method	**train**	**test**	**train**	**test**	**train**	**test**	**train**	**test**	**train**	**test**	**train**	**test**	**train**	**test**	**train**	**test**
Case V	AIM-rule	0.668	0.677	-	-	-	-	-	-	0.665	0.669	-	-	-	-	-	-
s+1=1	seq-BT	0.694	0.698	-	-	-	-	-	-	0.697	0.702	-	-	-	-	-	-
	PRIM	0.582	0.600	-	-	-	-	-	-	0.563	0.574	-	-	-	-	-	-
	MOB	0.984	0.983	-	-	-	-	-	-	0.983	0.983	-	-	-	-	-	-
	PC-TSMCD	1	1	-	-	-	-	-	-	1	1	-	-	-	-	-	-
Case VI	AIM-rule	0.365	0.366	0.243	0.265	0.745	0.742	-	-	0.36	0.364	0.235	0.253	0.732	0.737	-	-
s+1=3	seq-BT	0.39	0.391	0.355	0.375	0.724	0.722	-	-	0.391	0.392	0.37	0.381	0.708	0.712	-	-
	PRIM	0.459	0.456	0.561	0.573	0.741	0.738	-	-	0.452	0.449	0.545	0.542	0.731	0.733	-	-
	MOB	0.471	0.465	0.626	0.609	0.456	0.45	0.341	0.341	0.399	0.393	0.518	0.504	0.484	0.478	0.166	0.169
	PC-TSMCD	0.966	0.915	0.983	0.944	0.959	0.897	0.96	0.921	0.947	0.907	0.965	0.934	0.941	0.893	0.938	0.906
	TSMCD	0.993	0.985	0.997	0.988	0.991	0.984	0.992	0.983	0.963	0.958	0.972	0.966	0.959	0.954	0.96	0.953
Case VII	AIM-rule	0.69	0.666	0.439	0.416	0.943	0.916	-	-	0.692	0.675	0.453	0.435	0.931	0.912	-	-
s+1=2	seq-BT	0.706	0.688	0.5	0.479	0.915	0.897	-	-	0.702	0.689	0.495	0.477	0.911	0.898	-	-
	PRIM	0.791	0.768	0.728	0.697	0.859	0.84	-	-	0.796	0.78	0.718	0.693	0.877	0.865	-	-
	MOB	0.7097	0.7021	0.7525	0.745	0.66	0.653	-	-	0.892	0.889	0.732	0.722	0.732	0.723	-	-
	PC-TSMCD	0.8796	0.8083	0.9205	0.8515	0.8346	0.7611	-	-	0.918	0.864	0.926	0.873	0.91	0.854	-	-
	TSMCD	0.998	0.989	0.998	0.989	0.997	0.988	-	-	0.999	0.993	0.999	0.994	0.999	0.993	-	-
Case VIII	AIM-rule	0.676	0.68	0.461	0.463	0.891	0.891	-	-	0.692	0.68	0.474	0.46	0.912	0.9	-	-
s+1=2	seq-BT	0.681	0.684	0.47	0.471	0.896	0.894	-	-	0.687	0.683	0.462	0.456	0.914	0.91	-	-
	PRIM	0.622	0.616	0.756	0.755	0.481	0.477	-	-	0.617	0.615	0.757	0.75	0.472	0.48	-	-
	MOB	0.737	0.723	0.734	0.717	0.74	0.729	-	-	0.506	0.499	0.596	0.581	0.418	0.417	-	-
	PC-TSMCD	0.85	0.827	0.902	0.878	0.796	0.781	-	-	0.977	0.949	0.987	0.961	0.966	0.94	-	-
	TSMCD	0.991	0.982	0.991	0.983	0.99	0.982	-	-	0.999	0.994	0.999	0.994	0.999	0.993	-	-

For cases V to VIII, we notice that PC-TSMCD method outperforms AIM-rule, seq-BATTing, PRIM, and MOB methods with smaller prediction error. Moreover, our methods can correctly identify the number of subgroups, whereas the number of subgroups given by MOB is much larger than 3 in case VI. We consider a relatively high-dimensional data set in case VII with $p = 50$. When the number of subjects in a subgroup is smaller than 50, MOB cannot work any more because the underlying regression model cannot be fit with large p and small n. While AIM-rule, seq-BATTing, PRIM, and PC-TSMCD can still apply, we need additional high-dimensional model fitting program such as `glmnet` function in the software `R`.

We also consider the treatment effect to mimic a predictive signature development scenario in cases IV and VIII. Our proposed methods perform better than other compared approaches in terms of smaller mean square error for both training sample and test sample. We also observe from Table 2 that the rate of true subgroup numbers identified is very satisfactory. This suggests that our methods can be potentially applied to predictive signature development studies.

5 | APPLICATIONS

In this section, we apply our method to a clinical trial for Scleroderma patients and a breast cancer study (BCS). Both data are available from the author in this paper.

5.1 | Bovine Collagen Clinical Trial (BCCT)

We apply our subgrouping methods to an NIH-sponsored randomized Bovine Collagen Trial for Scleroderma patients at 12 centers in the USA.[55,56] Patients with diffuse Scleroderma were enroled in this multicenter phase II double-blind placebo controlled trial and a total of 831 observations were collected. Patients were randomized to receive oral native collagen at a dose of 500 μg/day or a similar appearing placebo. The Modified Rodnan Skin Score (MRSS) was the primary outcome variable and other key variables were disability index of the Health Assessment Questionnaire (HAQ), patient's global assessment, patients pain assessment, and physicians global assessment. To implement the proposed methods to predict MRSS, we consider six predictor variables $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_6$: over (disease progression), pain (index of pain), haq (health

TABLE 3 Results for BCCT data. MSE.tr(sd) and MSE.te(sd): averaged mean squared error among 100 iterations of 5-fold cross validation for training sample and test sample; $\hat{s}+1$: mean of estimated number of subgroups; AIM-rule: multiplicative rules-based modification of the adaptive index model; seq-BT: sequential BATTing; PRIM: patient rule induction method; MOB: model-based recursive partitioning method; 1-TSMCD: TSMCD with covariates as thresholding variable; F1-TSMCD: TSMCD with factor analysis (weighted least square estimator); F2-TSMCD: TSMCD with factor analysis (Thomson's regression estimator); PC-TSMCD: TSMCD with principal components; LM: linear regression model

Method	MSE.tr(sd)	MSE.te(sd)	$\hat{s}+1$ (sd)
AIM-rule	0.7794(0.0093)	0.8347(0.0170)	2(0)
seq-BT	0.7725(0.0116)	0.8342(0.0169)	2(0)
PRIM	0.7793(0.0102)	0.8312(0.0151)	2(0)
MOB	0.6977(0.0096)	0.8535(0.0577)	1.9020(0.1463)
1-TSMCD	0.7657(0.0060)	0.8299(0.0199)	1.3587(0.0769)
A-TSMCD	0.7753(0.0119)	0.8396(0.0176)	1.2140(0.1891)
F1-TSMCD	0.7438(0.0154)	0.8236(0.0201)	1.5840(0.2196)
F2-TSMCD	0.7447(0.0140)	0.8228(0.0220)	1.5520(0.1888)
PC-TSMCD	0.7592(0.0123)	0.8330(0.0133)	1.4000(0.2010)
LM	0.7610(0.0013)	0.8102(0.0121)	1(0)

assessment questionnaire), pga (patient self assessment of disease progression), dlcop (lung performance measurement), and age. After removing missing values, we have a sample of 295 observations in the downstream analysis. All variables are standardized with mean zero and unit variance.

We implement our proposed TSMCD methods along with AIM-rule, seq-BATTing, PRIM, MOB, and linear regression model (LM) to analyze the data. To evaluate the performance, we drive a 5-fold cross validation to this data set and repeat this procedure 100 times. The analysis results are summarized in Table 3. We also present box plots of MSE for both training and test sample in Figure 7. When we check the value of MSE, we can easily find that MOB has the smallest MSE for training sample, but for test sample, MOB has the largest prediction error with a large standard deviation. For training sample, the MSE of all subgrouping methods, expect MOB, is very close to the MSE of linear regression. For test sample, linear regression has the smallest MSE and MSE for other methods are also very close to it.

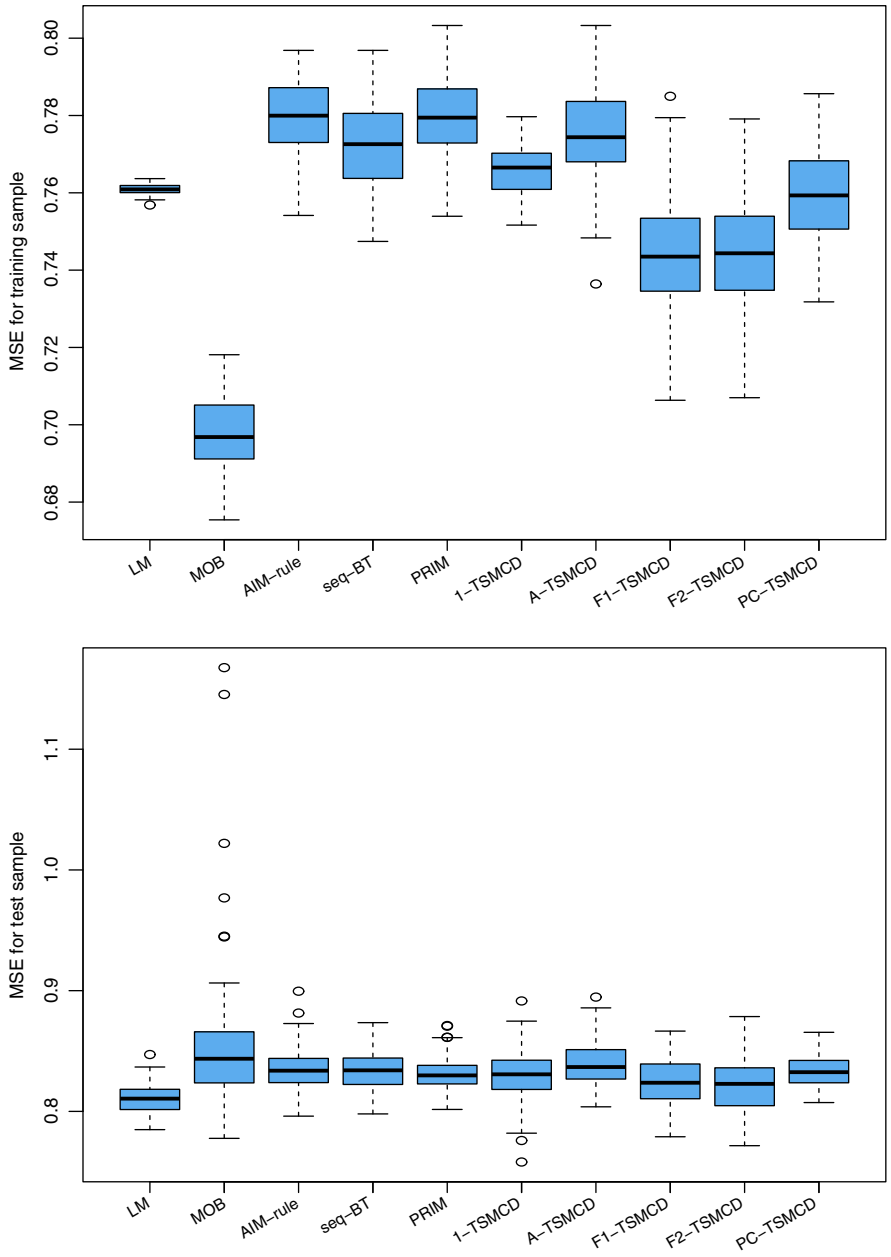

FIGURE 7 Box plots of mean squared error (MSE) for the training sample and test sample of Bovine Collagen Clinical Trial (BCCT) data [Colour figure can be viewed at wileyonlinelibrary.com]

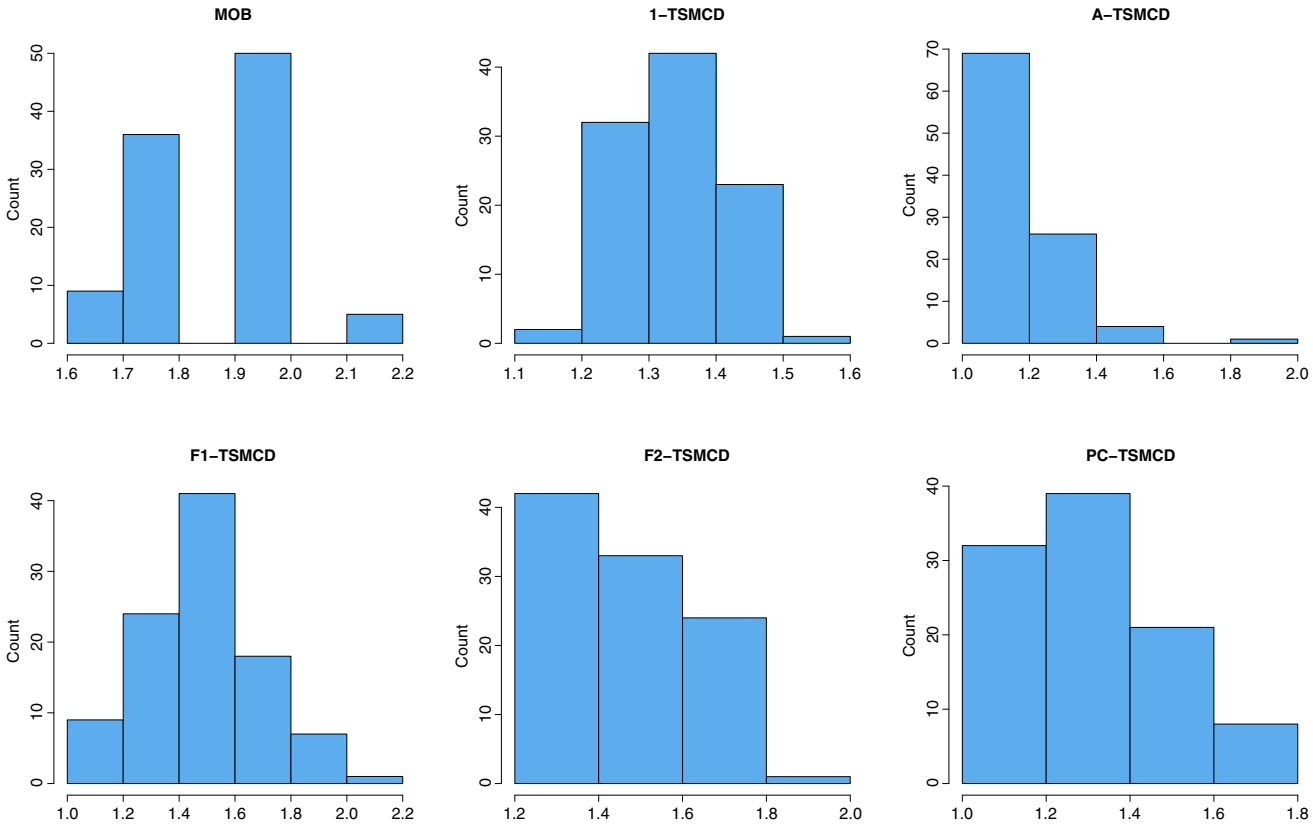

FIGURE 8 The frequency of subgroup numbers for Bovine Collagen Clinical Trial (BCCT) data [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 4 Results for the first procedure of BCS data. MSE(sd): averaged mean squared error for response(standard deviation); $\hat{s}+1$(sd): averaged estimated number of subgroups(standard deviation); AIM-rule: multiplicative rules-based modification of the adaptive index model; seq-BT: sequential BATTing method; PRIM: patient rule induction method; 1-TSMCD: TSMCD with covariates as thresholding variable; A-TSMCD: TSMCD with average of all covariates as thresholding variable; F1-TSMCD: TSMCD with factor analysis (weighted least square estimator); F2-TSMCD: TSMCD with factor analysis (Thomson's regression estimator); PC-TSMCD: TSMCD with principal components; LM: linear regression model; time: the mean computing time for each method

Method	MSE(sd)	$\hat{s}+1$(sd)	time(second)
AIM-rule	0.6763(0.0622)	2(0)	62.50
seq-BT	0.6916(0.0636)	2(0)	61.11
PRIM	0.6816(0.0732)	2(0)	18.79
1-TSMCD	0.7618(0.1042)	2.1314(0.5868)	249.21
A-TSMCD	0.7847(0.1536)	2.243(1.3715)	12.24
F1-TSMCD	0.6565(0.1190)	3.6700(1.2977)	75.82
F2-TSMCD	0.6762(0.1468)	3.0460(1.2518)	76.91
PC-TSMCD	0.6370(0.1523)	3.2900(1.3067)	57.87
LM	0.7384(0.0608)	1(0)	-

Table 3 shows the averaged numbers of subgroups identified by various methods and we observe that the AIM-rule, seq-BT, PRIM, and MOB give two subgroups and our proposed methods give about 1.4 subgroups. Figure 8 displays the frequency of identified number of subgroups except those for the AIM-rule, seq-BATTing, and PRIM methods because they always give two subgroups. Eyeballing the plots, we observe that A-TSMCD and F2-TSMCD prefer to take the sample

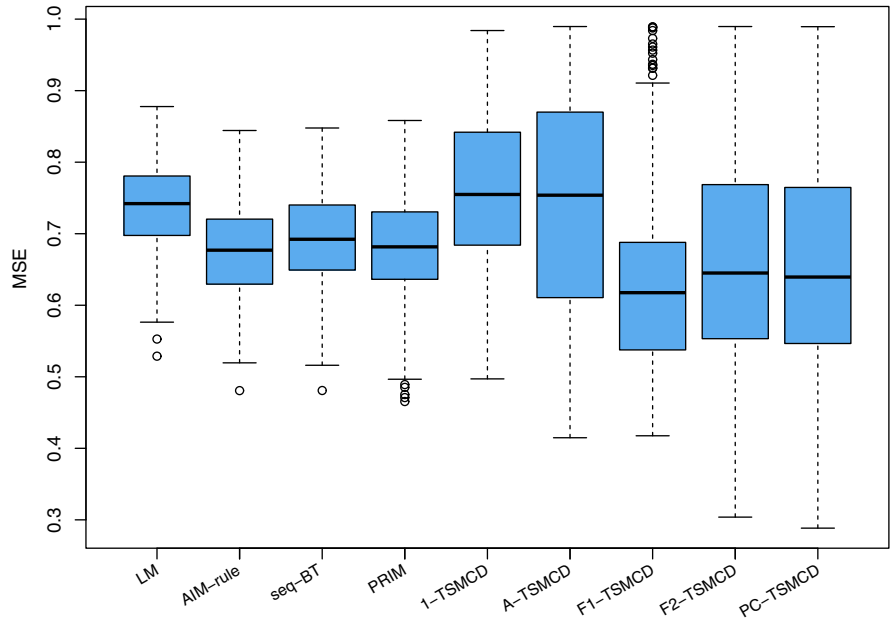

FIGURE 9 Box plots of mean squared error (MSE) for the first analysis procedure of breast cancer study (BCS) [Colour figure can be viewed at wileyonlinelibrary.com]

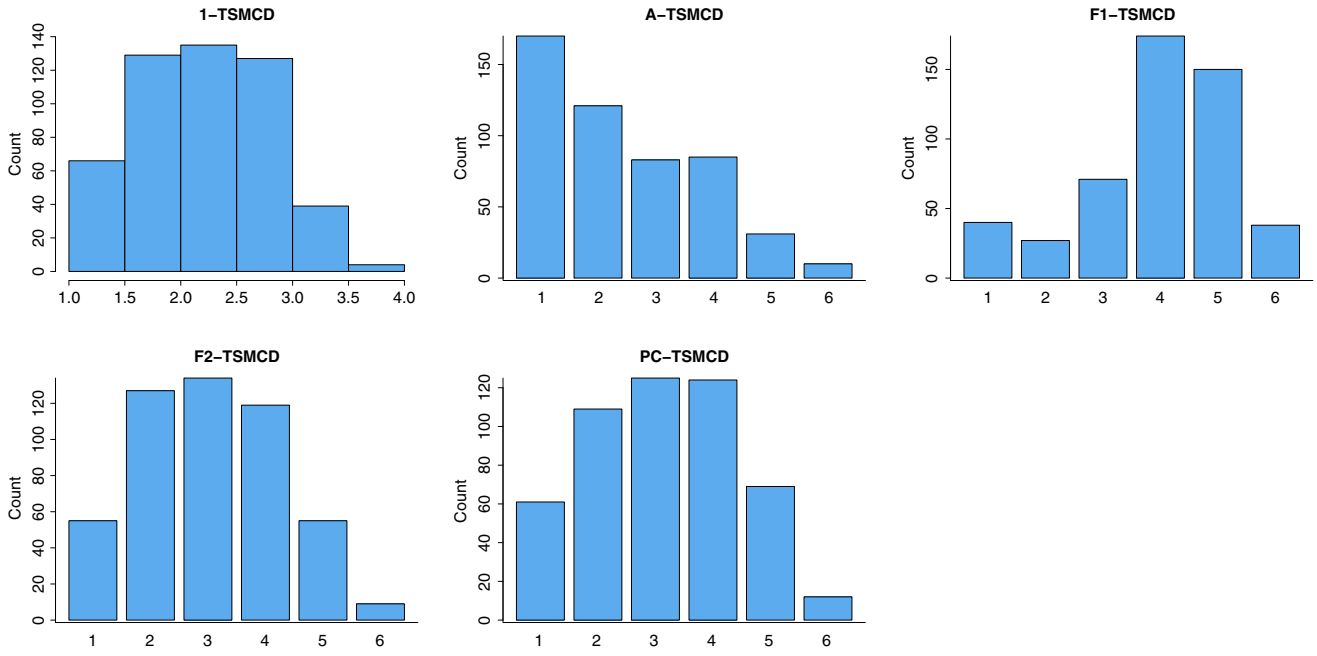

FIGURE 10 The frequency of the number of subgroups for the first analysis procedure of Breast Cancer Study (BCS). TSMCD, two-stage multiple change-point detection [Colour figure can be viewed at wileyonlinelibrary.com]

as one group and MOB has a high chance to identify two subgroups. We conclude that this BCCT data set can be deemed as one group and we do not have to identify subgroups.

5.2 | Breast cancer study

We next apply our methods to a high-dimensional BCS first reported in the work of van 't Veer et al.[57] There were 97 lymph node-negative breast cancer patients who were 55 years old or younger in this study. Among them, 46 developed distant metastases within 5 years (metastatic outcome coded as 1) and 51 remained metastases free for at least 5 years (metastatic outcome coded as 0). Clinical risk factors (confounders) were age, tumor size, histological grade, angioinvasion, lymphocytic infiltration, estrogen receptor (ER), and progesterone receptor (PR) status. Expression levels for 24 481 gene probes were collected. Directly analyzing $>$ 20 000 genes was problematic and, after removing genes with severe missingness, we analyzed 24 188 genes. In this section, we implement three of our computational procedures to analyze this breast cancer data set.

In the first procedure, we apply a bootstrap approach to build predictive models. We randomly pick 20 genes as the covariates and use tumor size as the response variable. We implement our proposed methods to this bootstrap data set and repeat the random sampling and computation for 500 times.

Table 4 summarizes the results. For the 1-TSMCD method, we take the average of MSE, which is obtained by taking each of the 20 covariates as thresholding variable. Among all the methods, PC-TSMCD achieves the smallest prediction error. Using PC-based thresholding and factor-based thresholding lead to, on average, more than three subgroups, whereas other subgrouping methods only divide the sample into two groups. The PRIM method is fastest among all the subgrouping methods. Figure 9 displays the box plots of the MSE over 500 repetitions for all methods. For comparison, we also consider linear regression in this example. This approach does not yield a subgroup and has relatively higher prediction errors. Figure 10 displays the corresponding frequencies of the number of subgroups by the TSMCD methods. A-TSMCD treats all subjects as one group, whereas F2-TSMCD and PC-TSMCD tend to split the sample into 2, 3, and 4 subgroups.

Yu et al[58] performed a screening analysis on this breast cancer data using the receiver operating characteristic–based approach by adjusting for the clinical risk factors. Their methods produced 10 important genes and 10 genes by practical ranking. In this second analysis procedure, we use the same 20 genes and build models using them as predictors

TABLE 5 Results for the second procedure of BCS data. MSE.tr(sd) and MSE.te(sd): averaged mean squared error among 100 iterations of 2-fold cross validation for training sample and test sample; $\hat{s}+1$: mean of estimated number of subgroups; AIM-rule: multiplicative rules-based modification of the adaptive index model; seq-BT: sequential BATTing; PRIM: patient rule induction method; MOB: model-based recursive partitioning method; 1-TSMCD: TSMCD with covariates as thresholding variable; F1-TSMCD: TSMCD with factor analysis (weighted least square estimator); F2-TSMCD: TSMCD with factor analysis (Thomson's regression estimator); PC-TSMCD: TSMCD with principal components; LM: linear regression model

Method	MSE.tr(sd)	MSE.te(sd)	$\hat{s}+1$ (sd)
AIM-rule	0.4989(0.1201)	1.4622(0.3930)	2(0)
seq-BT	0.5224(0.1317)	1.5061(0.4940)	2(0)
PRIM	0.5259(0.1256)	1.4649(0.4564)	2(0)
1-TSMCD	0.3817(0.0422)	1.9609(0.3380)	2.5194(0.1370)
A-TSMCD	0.4138(0.1271)	2.0159(0.6202)	2.4280(0.5302)
F1-TSMCD	0.4374(0.0813)	1.3328(0.2089)	2.4440(0.4174)
F2-TSMCD	0.4401(0.0859)	1.3682(0.2392)	2.4520(0.4685)
PC-TSMCD	0.4424(0.0927)	1.3659(0.2768)	2.4920(0.4622)
LM	0.4481(0.0469)	2.5049(0.8409)	1(0)

$\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{20}$. The 20 genes numbers are 10755, 16274, 13143, 10513, 19642, 7374, 22328, 296, 11285, 4682, 271, 403, 8, 272, 1439, 24023, 921, 194, 23488, and 593.

We use a 2-fold cross validation with 100 repetitions on this data. Table 5 reports the MSE results for the various models. We observe that, although 1-TSMCD has the smallest MSE for training sample, it makes a bad prediction on the test sample. F1-TSMCD, F2-TSMCD, and PC-TSMCD achieve the smaller MSE for both the training sample and test sample and they produce, on average, more than two subgroups. Figure 11 compares MSEs for these methods and Figure 12 displays the bar plot for each method to compare the frequency of the number of subgroups. Overall, our proposed approaches split the sample into more than two subgroups for this data set.

In the third analysis, we implement our proposed methods to analyze the genes identified by Cheng et al,[59] who analyzed the same data using a forward variable selection method. Four genes, namely, gene 2098, 23300, 19846, and 7844, were

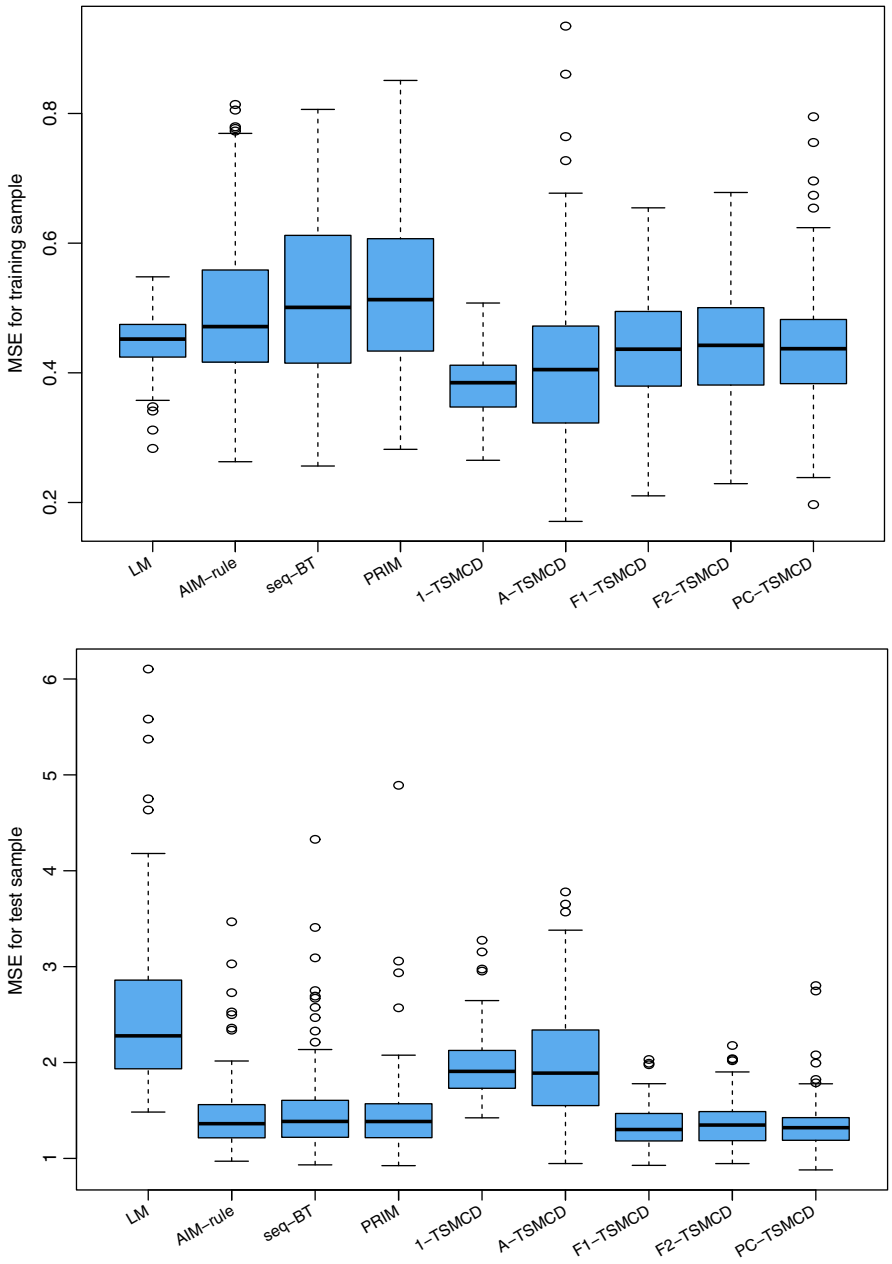

FIGURE 11 Box plots of mean squared error (MSE) for the training sample and test sample of the second procedure for Breast Cancer Study (BCS) [Colour figure can be viewed at wileyonlinelibrary.com]

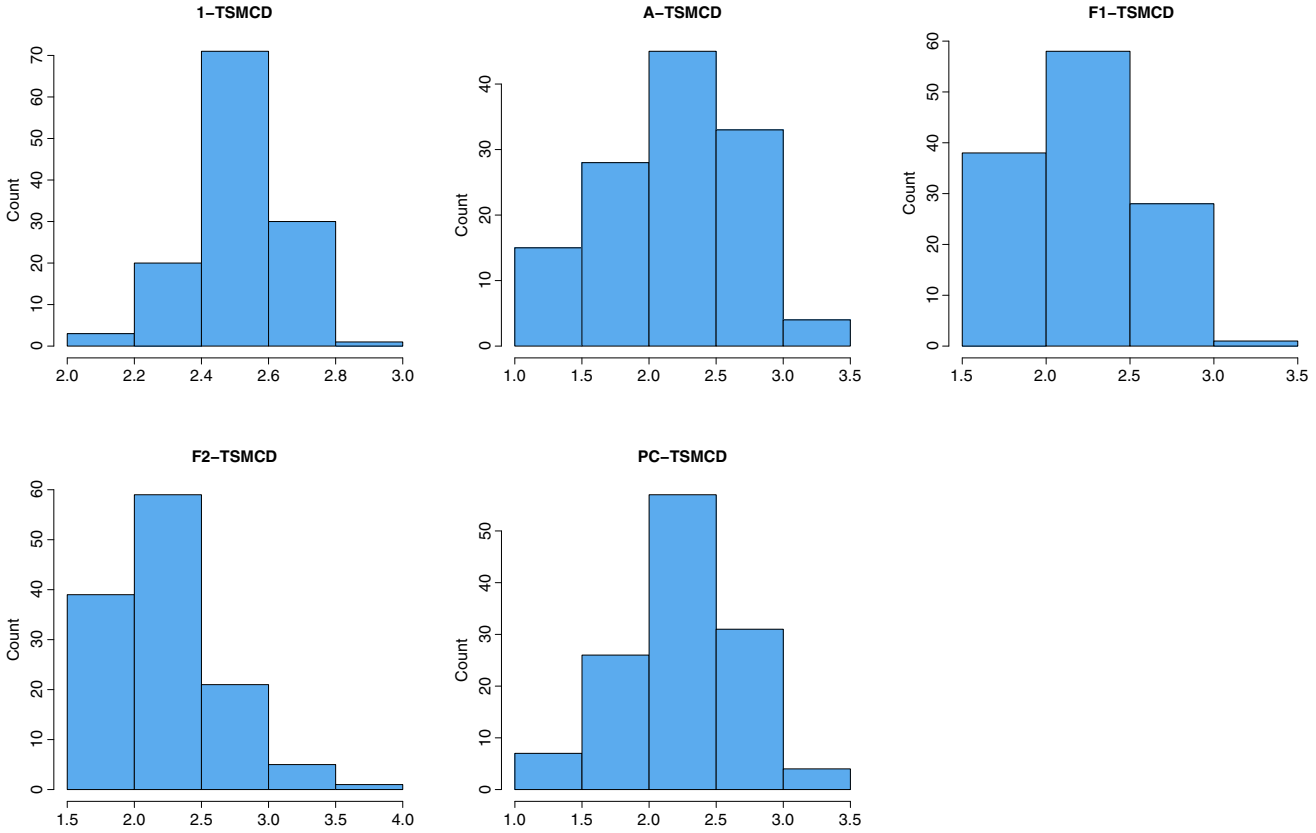

FIGURE 12 The frequency of the number of subgroups for the second procedure for Breast Cancer Study (BCS). TSMCD, two-stage multiple change-point detection [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 6 Results for the third procedure of BCS data. MSE.tr(sd) and MSE.te(sd): averaged mean squared error among 100 iterations of 2-fold cross validation for training sample and test sample; $\hat{s}+1$: mean of estimated number of subgroups; AIM-rule: multiplicative rules-based modification of the adaptive index model; seq-BT: sequential BATTing; PRIM: patient rule induction method; MOB: model-based recursive partitioning method; 1-TSMCD: TSMCD with covariates as thresholding variable; F1-TSMCD: TSMCD with factor analysis (weighted least square estimator); F2-TSMCD: TSMCD with factor analysis (Thomson's regression estimator); PC-TSMCD: TSMCD with principal components; LM: linear regression model

Method	MSE.tr(sd)	MSE.te(sd)	$\hat{s}+1$ (sd)
AIM-rule	0.5988(0.0511)	1.0308(0.1677)	2(0)
seq-BT	0.6114(0.0551)	1.0290(0.1498)	2(0)
PRIM	0.6050(0.0591)	1.0363(0.1579)	2(0)
MOB	0.6733(0.0302)	0.9109(0.1302)	1.0050(0.05)
1-TSMCD	0.6352(0.0457)	1.1055(0.1928)	1.3350(0.1904)
A-TSMCD	0.6553(0.0868)	1.3230(1.1757)	1.2450(0.3138)
F1-TSMCD	0.6236(0.0898)	1.1850(0.9451)	1.3750(0.3718)
F2-TSMCD	0.6236(0.0898)	1.1850(0.9451)	1.3750(0.3718)
PC-TSMCD	0.6220(0.065)	0.9859 (0.1434)	1.3950(0.3356)
LM	0.6742(0.0285)	0.9105(0.1301)	1(0)

identified and we used them as $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_4$ in our models. As in the second analysis, we also drive the 2-fold cross validation over 100 iterations. Table 6 outlines the averaged mean square error and the averaged number of subgroups. In this case, MOB and LM have the largest MSEs, but these two methods achieve the smallest prediction errors for the test sample. Although comparing with LM and MOB other subgrouping approaches have relative smaller MSE for the training data set, they do report larger prediction errors for the test data set. Moreover, we also can see that the variance of our proposed methods, A-TSMCD, F1-TSMCD, F2-TSMCD, and PC-TSMCD, are larger than AIM-rule, Seq-BT, PRIM, LM and MOB in Figure 13. Figure 14 shows that almost all our proposed methods prefer to have one group for the data set.

The anonymized data sets used in this paper will be made available upon acceptance of the paper.

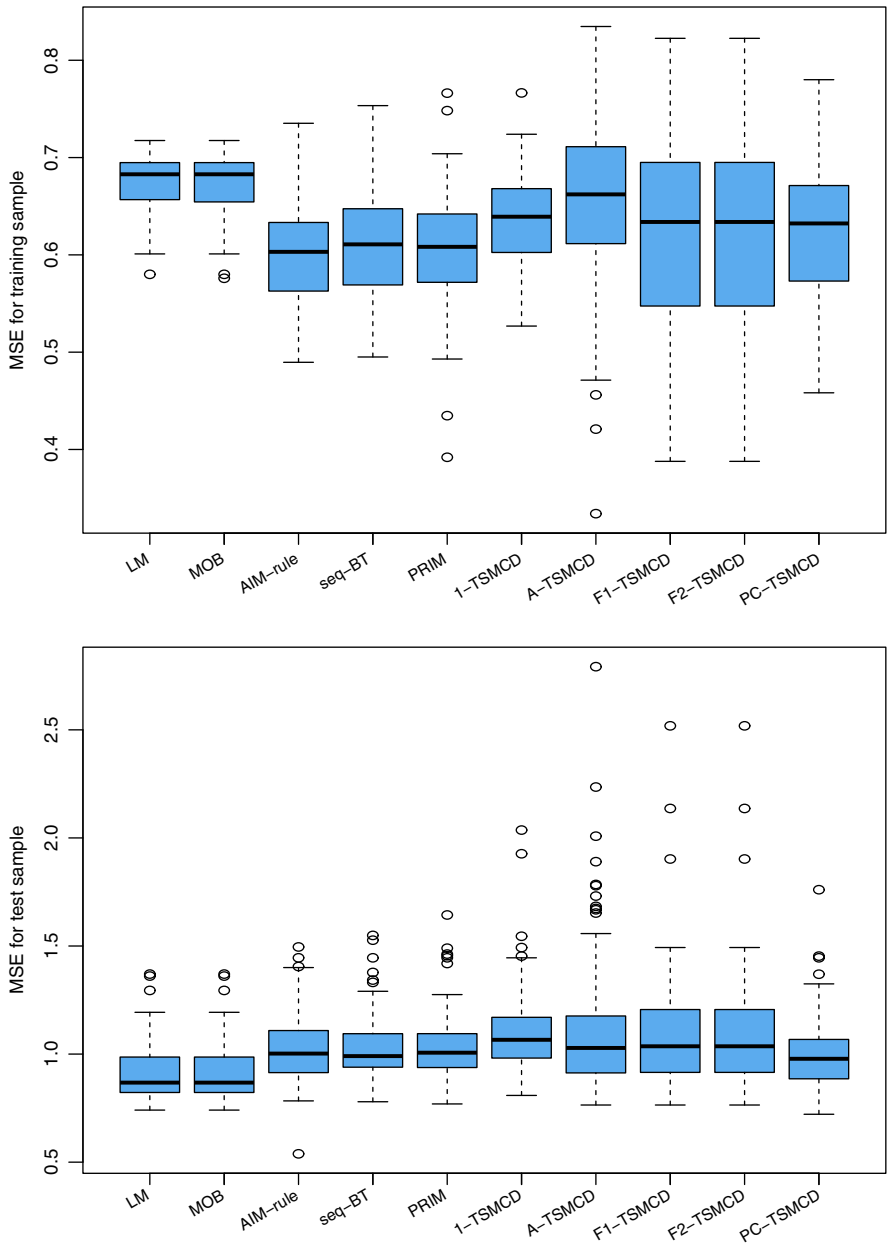

FIGURE 13 Box plots of mean squared error (MSE) for the training sample and test sample of the third procedure for Breast Cancer Study (BCS) [Colour figure can be viewed at wileyonlinelibrary.com]

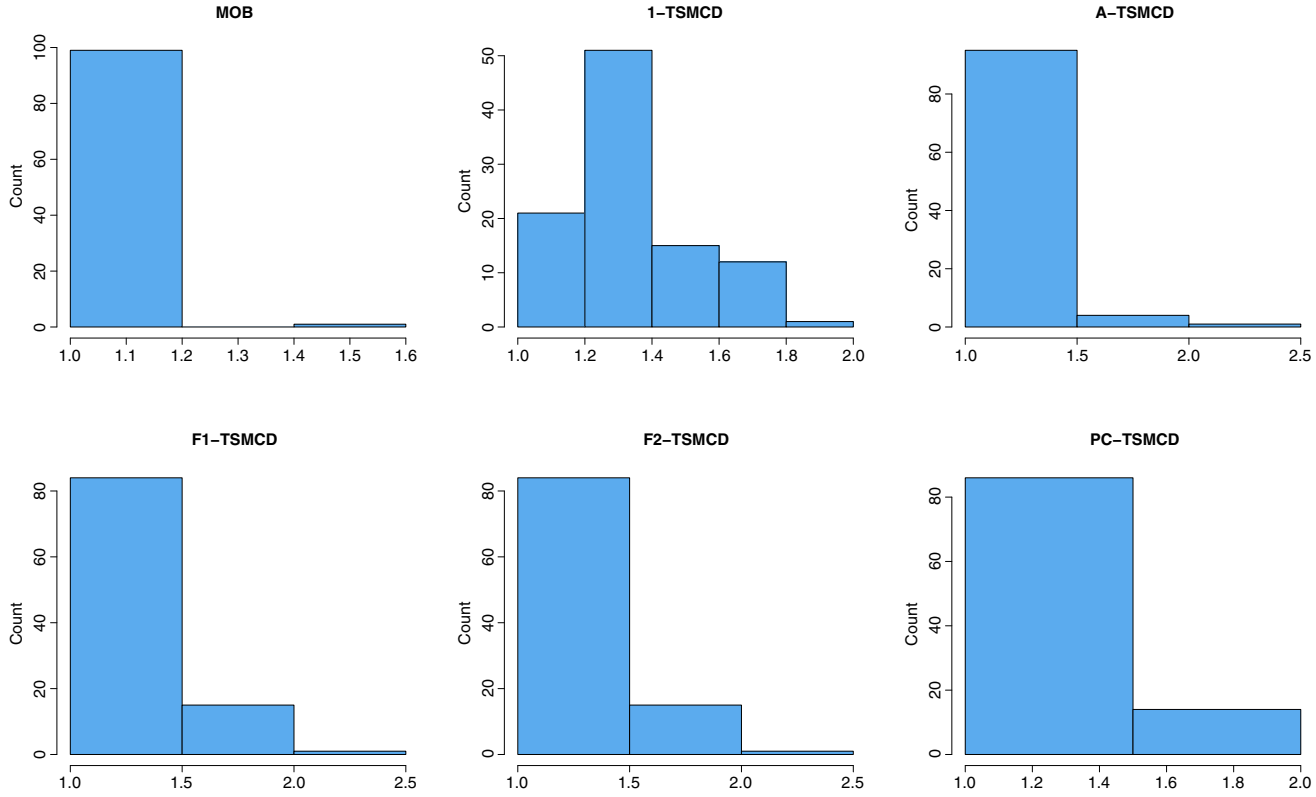

FIGURE 14 The frequency of the number of subgroups for the third procedure for Breast Cancer Study (BCS). MOB, model-based recursive partitioning; TSMCD, two-stage multiple change-point detection [Colour figure can be viewed at wileyonlinelibrary.com]

6 | DISCUSSION

Our work focuses on estimating the thresholding variable, which is an important component in the statistical analysis of precision medicine. We consider a variety of candidates for a thresholding variable and, via extensive simulations, provide recommendations for their choices. For example, when we have high-dimensional problems with a large number of covariates, thresholding variables from PC-based methods tend to yield more efficient results. On the other hand, in low-dimensional problems such as the bovine collagen trial data, factor analysis may yield quite helpful results and ought to be treated as a competitive approach for subgroup analysis.

We develop a general framework for constructing thresholding variables. The options can be based on a weighted sum of covariates, factors, principal components, or some other related variables as thresholding variables. This allows users to have more choices when they do with subgrouping problems. Our extensive numerical works suggest that these flexible choices may lead to more accurate prediction results for the medical outcome variables. When the analysts aim at improving the prediction value, it is especially appealing to consider our proposal. Unfortunately, it is sometimes not as easy to interpret the complicated thresholding variables as using a single covariate, which is the price we have to pay for this general thresholding variable.

Furthermore, our proposed methods can split the groups into more than two subgroups, whereas some traditional methods like AIM-rule, seq-BATTing, and PRIM only generate two subgroups. Allowing more subgroups in general could discover more data heterogeneity and identify hidden subcategory in a mixing population. Such multigroup structure may be fairly plausible for data sets with large n and large p, which becomes more and more common recently.

Many tree-based methods order the thresholding variables by Gini index, entropy function, or information gain.[4,60] SIDES[11] finds the best five (default) candidate covariates by optimizing some desired criteria (p-value, treatment effect, safety, and so on). AIM-rule and sequential BATTing are based on score test statistics. Our proposed methods are different from these recursive partitioning methods because we do not refit the change point models. Because we assume a true data generating model in this paper, the subgroup identification problem becomes a model estimation problem. Once the

model parameters are estimated, we obtain the grouping results automatically. Our procedure thus distinguishes from the machine learning procedures, where data are repeatedly analyzed to reinforce the final prediction performance.

In this paper, we mainly focus on continuous outcome with a Gaussian error distribution. Our proposal can be extended to address nonnormally distributed response variables too. For example, we may consider the censored survival time problem, in which case we can insert the Kaplan-Meier weights in the least squares to deal with the random censoring as in the work of Li and Jin.[19] Most existing subgroup methods can also work for dichotomous or multicategory outcomes, under a likelihood estimation framework. We then need to modify the objective function in this paper and seek a penalized maximum likelihood estimator. More theoretical and numerical works are under development.

ACKNOWLEDGEMENTS

The authors thank the two referees for their helpful comments that substantially improve the original submission. The work was partly supported by the Academic Research Funds R-155-000-205-114 and R-155-000-195-114 and Tier 2 MOE funds in Singapore MOE2017-T2-2-082: R-155-000-197-112 (Direct cost) and R-155-000-197-113 (IRC). The research of Wong was partially supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R01GM107639.

ORCID

Jialiang Li https://orcid.org/0000-0002-9704-4135
Weng Kee Wong https://orcid.org/0000-0001-5568-3054

REFERENCES

1. Clark GM, Zborowski DM, Culbertson JL, et al. Clinical utility of epidermal growth factor receptor expression for selecting patients with advanced non-small cell lung cancer for treatment with erlotinib. *J Thorac Oncol*. 2006;1:837-846.
2. Morgan JN, Sonquist JA. Problems in the analysis of survey data, and a proposal. *J Am Stat Assoc*. 1963;58:415-434.
3. Messenger R, Mandell L. A modal search technique for predictive nominal scale multivariate analysis. *J Am Stat Assoc*. 1972;67:768-772.
4. Breiman L, Friedman J, Olshen R, Stone C. *Classification and Regression Trees*. Boca Raton, FL: CRC Press; 1984.
5. Zeileis A, Hothorn T, Hornik K. Model-based recursive partitioning. *J Comput Graph Stat*. 2008;17(2):492-514.
6. Loh WY. Fifty years of classification and regression trees. *Int Stat Rev*. 2014;82(3):329-348.
7. Loh WY. Classification and regression trees. *Wiley Interdiscip Rev Data Min Knowl Discov*. 2011;1:14-23.
8. Loh WY. Regression trees with unbiased variable selection and interaction detection. *Stat Sin*. 2002;12:361-386.
9. Loh WY, He X, Man M. A regression tree approach to identifying subgroups with differential treatment effects. *Statist Med*. 2015;34:1818-1833.
10. Su X, Tsai CL, Wang H, Nickerson DM, Li B. Subgroup analysis via recursive partitioning. *J Mach Learn Res*. 2009;10:141-158.
11. Lipkovich I, Dmitrienko A, Denne J, Enas G. Subgroup identification based on differential effect search - a recursive partitioning method for establishing response to treatment in patient subpopulations. *Statist Med*. 2011;30:2601-2621.
12. Foster JC, Taylor JM, Ruberg SJ. Subgroup identification from randomized clinical trial data. *Statist Med*. 2011;30:2867-2880.
13. Laber EB, Zhao YQ. Tree-based methods for individualized treatment regimes. *Biometrika*. 2015;102(3):501-514.
14. Fu H, Zhou J, Faries DE. Estimating optimal treatment regimes via subgroup identification in randomized control trials and observational studies. *Statist Med*. 2016;35:3285-3302.
15. Doove LL, Dusseldorp E, Van Deun K, Van Mechelen I. A comparison of five recursive partitioning methods to find person subgroups involved in meaningful treatment—subgroup interactions. *Adv Data Anal Classif*. 2014;8(4):403-425.
16. Lipkovich I, Dmitrienko A, D'Agostino RB. Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Statist med*. 2017;36:136-196.
17. Bai J. Common breaks in means and variances for panel data. *J Econ*. 2010;157:78-92.
18. Qian J, Su L. Shrinkage estimation of common breaks in panel data models via adaptive group fused Lasso. *J Econ*. 2016;191:86-109.
19. Li J, Jin B. Multi-threshold accelerate failure time model. *Ann Stat*. 2018;46:2657-2682.
20. Huang X, Sun Y, Trow P, et al. Patient subgroup identification for clinical drug development. *Statist Med*. 2017;36:1414-1428.
21. Chen S, Tian L, Cai T, Yu M. A general statistical framework for subgroup identification and comparative treatment scoring. *Biometrics*. 2017;73(4):1199-1209.
22. Schnell PM, Tang Q, Offen WW, Carlin BP. A Bayesian credible subgroups approach to identifying patient subgroups with positive treatment effects. *Biometrics*. 2016;72:1026-1036.
23. Simon R. Bayesian subset analysis: application to studying treatment-by-gender interactions. *Statist Med*. 2002;21:2909-2916.

24. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York, NY: Springer-Verlag New York; 2009.
25. Bai J. Inferential theory for factor models of large dimensions. *Econometrica*. 2003;71:135-171.
26. Kustra R, Shioda R, Zhu M. A factor analysis model for functional genomics. *BMC Bioinformatics*. 2006;7:216.
27. Pournara I, Wernish L. Factor analysis for gene regulatory networks and transcription factor activity profiles. *BMC Bioinformatics*. 2007;8:61.
28. Friguet C, Kloareg M, Causeur D. A factor model approach to multiple testing under dependence. *J Amer Stat Assoc*. 2009;104:1406-1415.
29. Desai KH, Storey JD. Cross-dimensional inference of dependent high-dimensional data. *J Amer Stat Assoc*. 2012;107:135-151.
30. Fan J, Ke Y, Sun Q, Zhou WX. Farm-Test:factor-adjusted robust multiple testing with approximate false discovery control. arXiv preprint arXiv:1711.05386. 2017.
31. Hotelling H. Analysis of a complex of statistical variables into principal components. *J Educ Psychol*. 1933;24(6):417-441.
32. Rao CR. The use and interpretation of principal component analysis in applied research. *Indian J Stat Ser A*. 1964;26(4):329-358.
33. Tian L, Tibshirani R. Adaptive index models for marker-based risk stratification. *Biostatistics*. 2011;12:68-86.
34. Chen G, Zhong H, Belousov A, Devanarayan V. A PRIM approach to predictive-signature development for patient stratification. *Statist Med*. 2015;34:317-342.
35. Friedman JH, Fisher NI. Bump hunting in high-dimensional data. *Stat Comput*. 1999;9:123-143.
36. LeBlanc M, Jacobson J, Crowley J. Partitioning and peeling for constructing prognostic groups. *Stat Methods Med Res*. 2002;11(3):247-274.
37. LeBlanc M, Moon J, Crowley J. Adaptive risk group refinement. *Biometrics*. 2005;61(2):370-378.
38. Polonik W, Wang Z. PRIM analysis. *J Multivar Anal*. 2010;101(3):525-540.
39. Anderson TW. *An Introduction to Multivariate Statistical Analysis*. 3rd ed. Hoboken, NJ: John Wiley and Sons Inc; 2003.
40. Giri NC. *Multivariate Statistical Analysis*. 2nd ed. New York, NY: Marcel Dekker; 1996.
41. Thomson GH. *The Factorial Analysis of Human Ability*. London, UK: London University Press; 1951.
42. Jolliffe IT. *Principal Component Analysis*. 2nd ed. New York, NY: Springer-Verlag New York; 2002.
43. Golub GH, Van Loan CF. *Matrix Computations*. 3rd ed. Baltimore, MD: Johns Hopkins University Press; 1996.
44. Jung S, Marron JS. PCA consistency in high dimension, low sample size context. *Ann Stat*. 2009;37:4104-4130.
45. Paul D. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Stat Sin*. 2007;17(4):1617-1642.
46. Zou H, Hastie T, Tibshirani R. Sparse principal components analysis. *J Comput Graph Stat*. 2006;15(2):265-286.
47. Bair E, Hastie T, Paul D, Tibshirani R. Prediction by supervised principal components. *J Amer Stat Assoc*. 2006;101(473):119-137.
48. Barshan E, Ghodsi A, Azimifar Z, Jahromi MZ. Supervised principal component analysis: visualization, classification and regression on subspaces and submanifolds. *Pattern Recogn*. 2011;44(7):1357-1371.
49. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Amer Stat Assoc*. 2001;96:1348-1360.
50. Zhang C. Nearly unbiased variable selection under minimax concave penalty. *Ann Stat*. 2010;38:894-942.
51. Jin B, Shi X, Wu Y. A novel and fast methodology for simultaneous multiple structural break estimation and variable selection for nonstationary time series models. *Stat Comput*. 2013;23:221-231.
52. Wang H, Li B, Leng C. Shrinkage tuning parameter selection with a diverging number of parameters. *J R Stat Soc Ser B Stat Methodol*. 2009;71(3):671-683.
53. Lee ER, Noh H, Park BU. Model Selection via Bayesian information criterion for quantile regression models. *J Amer Stat Assoc*. 2014;109(505):216-229.
54. Luo S, Chen Z. Extended BIC for linear regression models with diverging number of relevant features and high or ultra-high feature spaces. *J Stat Plan Inference*. 2013;143:494-504.
55. Postlethwaite AE, Wong WK, Clements P, et al. Oral collagen I (CI) in diffuse systemic sclerosis (SSc): I. Oral CI does not improve skin in all patients but may improve skin in late disease. *Arthritis Rheum*. 2008;58:1810-1822.
56. Li J, Wong WK. A semi-parametric analysis for identifying Scleroderma patients responsive to an anti-fibrotic agent. *Contemp Clin Trials*. 2009;30:105-113.
57. van 't Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*. 2002;415:530-536.
58. Yu T, Li J, Ma S. Adjusting confounders in ranking biomarkers: a model-based ROC approach. *Brief Bioinform*. 2012;13(5):513-523.
59. Cheng MY, Honda T, Zhang JT. Forward variable selection for sparse ultra-high dimensional varying coefficient models. *J Amer Stat Assoc*. 2016;111(515):1209-1221.
60. Quinlan JR. Induction of decision trees. *Mach learn*. 1986;1(1):81-106.

How to cite this article: Wang J, Li J, Li Y, Wong WK. A model-based multithreshold method for subgroup identification. *Statistics in Medicine*. 2019;38:2605–2631. https://doi.org/10.1002/sim.8136