GEORGETOWN UNIVERSITY

Yuan A, Chen XF, Zhou YZ and Tan M. Robust subgroup analysis with semiparametric model in precision medicine. Statistics Medicine (accepted). 2018. **Ming Tan <mtt34@georgetown.edu>**

Fwd: Statistics in Medicine - Accept - Manuscript SIM-17-0424.R2

1 message

Ao Yuan <ay312@georgetown.edu> Fri, Jan 26, 2018 at 2:05 PM
To: xiaofeic@smu.edu, G U <xc81@georgetown.edu>, Yizhao Zhou <yz459@georgetown.edu>, Ming Tan <mtt34@georgetown.edu>

---------- Forwarded message ----------
From: Statistics in Medicine <onbehalfof@manuscriptcentral.com>
Date: Fri, Jan 26, 2018 at 1:57 PM
Subject: Statistics in Medicine - Accept - Manuscript SIM-17-0424.R2
To: yuanao@hotmail.com, ay312@georgetown.edu

Date: 26-Jan-2018

Manuscript Number: SIM-17-0424.R2

Title: "Sub-group analysis with semiparametric model in precision medicine"

Dear Dr. Yuan:

I am pleased to inform you that your Research Article has been accepted for publication in Statistics in Medicine.

The publisher is able to access the final version of your Research Article online.

Your article cannot be published until you have signed the appropriate license agreement. Within the next few days you will receive an email from Wiley's Author Services system which will ask you to log in and will present you with the appropriate license for completion.

If your paper contains Supporting Information: Please note that materials submitted as Supporting Information are authorized for publication alongside the online version of the accepted paper. It is the responsibility of the authors to supply any necessary permissions to the editorial office.

If you feel that your article will have broad appeal beyond the readership of Statistics in Medicine, you may wish to consider providing a "layman's abstract" to appear on Wiley's statistics community website, Statistics Views. StatisticsViews.com attracts

approximately 20,000 visitors a month and has a substantial social media following on both Twitter and Facebook. A large proportion of StatisticsViews's visitors are students who find layman's abstracts of articles very useful in their education.

This abstract or commentary should be written in the third person, be substantially less technical than the formal abstract, and be at least 200 words. It should explain the importance of your work in a broader context, appealing to a non-specialist statistical audience. We may use this information to highlight your research on Statistics Views and our social media networks. When ready, you may email your layman's abstract to Alison Oliver (aoliver@wiley.com).

Should you have any questions about this process please contact the Journal Administrator, Suzanne Thompson (stompson@bu.edu).

Thank you for your support of Statistics in Medicine. We look forward to seeing more of your work in the future.

Sincerely,

Dr. Ralph D'Agostino Sr.
Editor
Statistics in Medicine

P.S. Bring your research to life by creating a video abstract for your article! Wiley partners with Research Square to offer a service of professionally produced video abstracts. Learn more about video abstracts at http://www.wileyauthors.com/videoabstracts and purchase one for your article at https://www.researchsquare.com/wiley/ or through your Author Services account. If you have any questions, please direct them to videoabstracts@wiley.com.

Subgroup analysis with semiparametric models in precision medicine

(Running title: Sub-group analysis with semiparametric model)

Ao Yuan, Xiaofei Chen, Yizhao Zhou, Ming T. Tan

Department of Biostatistics, Bioinformatics and Biomathematics,

Georgetown University, Washington DC 20057, USA

Correspondence: Ao Yuan, ay312@georgetown.edu, Ming T. Tan, mtt34@georgetown.edu

Abstract

In analyzing clinical trials, one important objective is to classify the patients into treatment-favorable and non-favorable subgroups. Existing parametric methods are not robust, and the commonly used classification rules ignore the fact that the implications of treatment-favorable and non favorable subgroups can be different. To address these issues, we propose a semiparametric model, incorporating both our knowledge and uncertainty about the true model. The Wald statistics is used to test the existence of subgroups, while the Neyman-Pearson rule to classify each subject. Asymptotic properties are derived, simulation studies are conducted to evaluate the performance of the method, and then method is used to analyze a real world trial data.

Keywords. Clinical trial, EM-algorithm, Neyman-Pearson classification, profile likelihood, semiparametric model, sub-group.

1. Introduction.

In clinical trials, it is quite plausible that some of the treatments are particularly effective for some subgroups of the population but less so for others. For example, patients with ER-negative tumours benefited substantially from chemotherapy,

while those with ER-positive tumours did not benefit[1]. In another study Sabine[2] reported that the efficacy and toxicity profiles for the highly active antiretroviral therapy vary markedly across different subgroups of HIV-infected patients. Bonetti, M. and Gelber, R. D. [3] reported different effects on different subgroups of patients. Thus identifying subgroups particularly effective for some treatments, if they exist, is of important practical significance[4].

There are various methods for subgroup analysis. Cai et al.[5] proposed a semi-parametric approach to estimate the treatment difference among subgroups using kernel smoothing. Shen and He[6] proposed structured logistic-normal mixture model for a univariate continuous response outcome to test if there exists two different subgroups. However, under the null hypothesis of no subgroups, parameters in their logistic specification is not identifiable, the standard likelihood ratio statistic does not apply, which substantially complicates the testing of the existence of subgroups. In addition, their classification error is not controlled at a pre-specified level. Rothmann et al.[7] discussed issues of treatment negative patients for subgroup testing and analysis. Friede et al.[8] proposed conditional error approach for predefined subgroup and the full population as co-primary analyses. Song and Chi[9] and Alosh and Huque[10] addressed issues for testing effects difference between a predefined subgroup and the whole group. Sivaganesan et al.[11] and Jones et al.[12] considered Bayesian models. Su et al.[13] and Lipkovich et al.[14] used recursive partitioning method to determine the heterogeneity of the treatment effect across sub-populations. For 0-1 valued case-control data along with covariates, the commonly used method is logistic regression with forward/backward procedure to select the significant covariates representing the subgroups. For finite discrete random responses with covariates, Foster et al.[15] combine the methods of random forest and virtual twins, to partition the covariates space into sub-spaces, and classify each subject into one of the subgroups if his/her

covariates fall into the corresponding sub-space.

Subgroup analysis generally follows three steps: (i) specifying a model for the observed data, and (ii) estimating the parameters in the model (if any); and testing the existence of subgroups; after justification of existence, (iii) classifying each subjects into one of the subgroups, based on his/her covariate profile. Approaches for analysis can be generally divided into parametric and nonparametric. If correctly specified, the parametric model is simple to use and efficient in parameter estimation, but it is not robust to model misspecifications.

On the other hand, the nonparametric model is robust, but is less efficient. In practice, no subjectively specified model is exactly true, but often we do have some knowledge about the scientific background and the model to some extent. Therefore, we propose a semiparametric model, in which the prior knowledge about the true parametric model is incorporated into the model along with a nonparametric component to reflect model uncertainty. The extent of the correctness of the postulated parametric component is estimated from the data itself.

The classical classification method uses the Bayesian rule to classify the subjects, which treats each subgroup equally. For our subgroup analysis, correct classification of the treatment favorable subgroup is of much more importance than that for the other subgroup, and it is desirable to control the mis-classification rate on this subgroup. Thus, we use the Neyman-Pearson classification rule, in which the mis-classification error on the treatment favorable subgroup is controlled by a pre-specified level, while that on the other subgroup is minimized.

Specifically, we propose a two-step procedure for subgroup analysis: (i) we use nonparametric mixture model with a mixing proportion to estimate the regression parameters and test the existence of subgroups via the Wald test; if existence of subgroup is confirmed statistically, then (ii), we use the Neyman-Pearson rule to

classify each subject, this rule depends only on the likelihood ratio, not the mixing proportion. Therefore, our method has several advantages over existing methods for its a) incorporating current knowledge of the model, b) robustness, c) ease of use, d) avoidance of covariate dependent specification of mixing proportion for model complication, and e) control of miss-classification error at a pre-specified level.

On the other hand, since the latent subgroup memberships are unknown, the proposed model becomes a semiparametric mixture model whose parameters is known to be difficult t estimate. We will utilize the profile likelihood approach by first using the Expectation-Maximization (EM) algorithm structure to obtain a 'complete data' non-mixture model and treating the latent memberships as missing data. We then treat the nonparametric component in our model as nuisance parameters and profile them out, the parameter of interest is estimated via the resulting profile likelihood.

The rest of this article is organized as follows. Model formulation, estimation, and asympotic properties are studied in three subsections respectively in Section 2. We evaluate the performance of the method with simulation studies and apply it to analyze a real world clinical trial data (Section 3). We conclude with a discussion (Section 4).

2. The proposed method.

The observed data is $(y_i, \mathbf{x}_i), i = 1, ..., n$, where $y_i \in R$ is the response and $\mathbf{x}_i = (x_{i1}, ..., x_{id})' \in R^d$ is the covariate of the i-th subject. Each subject i receives one of k treatments, the treatment label (known) is in one of the components of $\mathbf{x}_i$, and we assume that bigger value of the response corresponds to better treatment effects. The goals are to test, for each treatment, the existence of treatment favorable and non-favorable subgroups, if the existence is confirmed, establish a rule based on the data to classify the subjects into the subgroups, and classify the coming patients into

one of the subgroups based on the rule. We concentrate on the case of one treatment with two possible subgroups. For this, we need to specify a model for the observed data, estimate the parameters in the model, then test the null hypothesis of no subgroups, and if the null hypothesis is rejected then classify the subjects. Below we describe these procedures one by one.

2.1 Semiparametric model formulation. Let δ_i be the latent indicator whether subject i belongs to the treatment favorable subgroup ($\delta_i = 1$) or not ($\delta_i = 0$). We specify the model as

$$y_i = \boldsymbol{\beta}'\mathbf{x}_i + \delta_i\eta + \epsilon_i, \qquad E(\epsilon_i) = 0, \tag{1}$$

where $\boldsymbol{\beta}$ is the unknown regression parameter vector, its first element represents the intercept, η is the additional effect of the treatment on the favorable sub-group, the constraint $\eta \geq 0$ is component-wise, and is used for the identifiability with the intercept vector term in $\boldsymbol{\beta}$. When intercept terms are included, we set $x_{i1} = 1$ for all i to incorporate them. Often we have some prior knowledge about the model or distribution of the error ϵ_i's, given by a known density function $g(\cdot)$, but not sure how accurate this model is. Thus conditioning on the status variable δ, we specify the distribution of ϵ_i's as a semiparametric geometric mixture

$$f(\epsilon|\delta) = \Delta^{-1} g^{\lambda}(\epsilon|\delta) h^{1-\lambda}(\epsilon|\delta), \qquad \Delta = \int g^{\lambda}(s) h^{1-\lambda}(s) ds, \tag{2}$$

where $g(\epsilon|\delta) = g(y - \boldsymbol{\beta}'\mathbf{x} - \delta\eta)$, $h(\epsilon|\delta) = h(y - \boldsymbol{\beta}'\mathbf{x} - \delta\eta)$, $g(\cdot)$ is the known density function, $h(\cdot)$ is unknown density function, and $0 < \lambda \leq 1$ is an unknown parameter representing our extent of belief on $g(\cdot)$. When $\lambda = 1$, the specified model g is correct, and when $\lambda \to 0$ the model is totally unknown and nonparametric and is to be estimated from the observed data. We assume that $h(\cdot)$ contains no portion of $g(\cdot)$, then model (2) is identifiable, in that if $\Delta_1^{-1} g^{\lambda_1}(\cdot) h_1^{1-\lambda_1}(\cdot) = \Delta^{-1} g^{\lambda}(\cdot) h^{1-\lambda}(\cdot)$, then

we must have $(\lambda_1, h_1(\cdot)) = (\lambda, h(\cdot))$. In fact, if $\Delta_1^{-1} g^{\lambda_1}(\cdot) h_1^{1-\lambda_1}(\cdot) = \Delta^{-1} g^{\lambda}(\cdot) h^{1-\lambda}(\cdot)$ (without loss of generality, assume $\lambda_1 > \lambda$), then we get $h^{1-\lambda}(\cdot) \propto g^{\lambda_1-\lambda} h_1^{1-\lambda_1}(\cdot)$, i.e., $h(\cdot)$ contains a proportion of $g(\cdot)$, which violates our assumption on $h(\cdot)$.

One advantage of the geometric mixture (2) over the arithmetic mixture $\lambda g(\epsilon|\delta) + (1-\lambda) h(\epsilon|\delta)$ is computational (see Section 2.5). The former has a log-likelihood $\lambda \log g(\epsilon|\delta) + (1-\lambda) \log h(\epsilon|\delta)$, while the log-likelihood for the latter is $\log\big(\lambda g(\epsilon|\delta) + (1-\lambda) h(\epsilon|\delta)\big)$. It is known that parameter estimation in the former log-likelihood is much easier than the latter. Olkin[16] considered an arithmetic mixture

$$\lambda g(\epsilon|\theta) + (1-\lambda) \hat{f}_n(\epsilon)$$

where $g(\cdot|\theta)$ is a given parametric density and $\hat{f}_n(\cdot)$ is the kernel density estimator. Also, model (2) can be interpreted as penalized likelihood which we will elaborate latter after model (4). Another advantage is its connection to the biased sampling model[17,18,19]. In that model, the density of the j-th biased subgroup is specified as $w_j(y) f(y) / \int w_j(t) f(t) dt$, where $w_j(\cdot)$ is often assumed known and $f(\cdot)$ is an unknown density. In this sense model (2) can be viewed as a geometric version of the biased sampling model, in which $g^{\lambda}(\cdot)$ represents our subjective guess (bias) or prior knowledge of the data model, and $h^{1-\lambda}(\cdot)$ represents the part unknown to us.

Let $\gamma = P(\delta = 1)$. More generally, $\gamma = \gamma(\mathbf{x})$ should be dependent on subject's covariates, such as parameterized by a logistic specification. However, it is known that its very difficult to estimate the parameters in mixing proportions. In Shen and He[6], they used covariate dependent mixing proportion and normal error distribution, their model loses the log-likelihood ratio test property and it not easy to use. For nonparametric error distribution, if covariate dependent mixing proportion is used, the model identifiability will be very challenging. Altstein and Li[20] used fixed mixing proportion for subgroup analysis using accelerated failure time model for censored

data. However, for the purpose of testing the existence of subgroups and estimation of regression parameters, only the mixture model with constant mixing proportion γ is needed.

As δ is latent, the density for $\epsilon = y - \boldsymbol{\beta}'\mathbf{x} - \delta\eta$ is the following mixture

$$f(\epsilon) = \gamma f(\epsilon|\delta = 1) + (1-\gamma) f(\epsilon|\delta = 0)$$

$$= \Delta^{-1}\bigg(\gamma g^{\lambda}(y - \boldsymbol{\beta}'\mathbf{x} - \eta)h^{1-\lambda}(y - \boldsymbol{\beta}'\mathbf{x} - \eta)\big) + (1-\gamma)g^{\lambda}(y - \boldsymbol{\beta}'\mathbf{x})h^{1-\lambda}(y - \boldsymbol{\beta}'\mathbf{x})\bigg). \quad (3)$$

Note that the proposed model has two levels of mixing, $f(\epsilon|\delta) = g^{\lambda}(y - \boldsymbol{\beta}'\mathbf{x} - \delta\eta)h^{1-\lambda}(y - \boldsymbol{\beta}'\mathbf{x} - \delta\eta)$ is a geometric mixture with the known $g(\cdot)$ and the unknown $h(\cdot)$; and $f(\epsilon) = \gamma f(\epsilon|\delta = 1) + (1-\gamma) f(\epsilon|\delta = 0)$ is another level of mixing over the two subgroups. The additive mixture is used for the latter to conform to the commonly used structure for subgroup analysis. So our model involves an additive mixture for distributions across the subgroups, arisen naturally due to the latent class variable, and within each subgroup, a geometric mixture of a known and an unknown density components.

To classify the subjects, we need first to estimate the parameters $\boldsymbol{\theta} := (\boldsymbol{\beta}, \eta, \lambda, \gamma)$ along with the infinite dimensional nuisance parameter $h(\cdot)$. However, it is known that estimation of parameters in mixture model (3) is not easy, and a common practice is to estimate the parameters in the corresponding 'complete' data model, with the latent data δ_i's added and treating them as missing.

The likelihood for the 'complete data' $(y_i, \mathbf{x}_i, \delta_i)$ $(i = 1, ..., n)$ is

$$L_n(\boldsymbol{\theta}, \Delta, h) = \Delta^{-n} \prod_{i=1}^{n} \bigg(\gamma h^{1-\lambda}(y_i - \boldsymbol{\beta}'\mathbf{x}_i - \eta) g^{\lambda}(y_i - \boldsymbol{\beta}'\mathbf{x}_i - \eta)\bigg)^{\delta_i}$$

$$\times \bigg((1-\gamma) h^{1-\lambda}(y_i - \boldsymbol{\beta}'\mathbf{x}_i) g^{\lambda}(y_i - \boldsymbol{\beta}'\mathbf{x}_i)\bigg)^{1-\delta_i},$$

the corresponding log-likelihood is

$$\ell_n(\boldsymbol{\theta}, \Delta, h) = \sum_{i=1}^{n} \delta_i \Big((1-\lambda) \log h(y_i - \boldsymbol{\beta}'\mathbf{x}_i - \eta) + \lambda \log g(y_i - \boldsymbol{\beta}'\mathbf{x}_i - \eta) + \log \gamma \Big)$$

$$+(1-\delta_i) \Big((1-\lambda) \log h(y_i - \boldsymbol{\beta}'\mathbf{x}_i) + \lambda \log g(y_i - \boldsymbol{\beta}'\mathbf{x}_i) + \log(1-\gamma) \Big) - n \log \Delta. \quad (4)$$

Model (4) has another interpretation. Let

$$D_n(h, g) = \textstyle\sum_{i=1}^{n} \sum_{\delta_i=0}^{1} \log \big(h(y_i - \boldsymbol{\beta}'\mathbf{x}_i - \delta_i \eta) / g(y_i - \boldsymbol{\beta}'\mathbf{x}_i - \delta_i \eta) \big) - n \log \Delta$$

be the Kullback-Leibler divergence between the two densities $h(\cdot)$ and $g(\cdot)$ based on the observed data. Then

$$\ell_n(\boldsymbol{\theta}, \Delta, h) = \sum_{i=1}^{n} \delta_i \Big(\log h(y_i - \boldsymbol{\beta}'\mathbf{x}_i - \eta) + \log \gamma \Big) +$$

$$+(1-\delta_i) \Big(\log h(y_i - \boldsymbol{\beta}'\mathbf{x}_i) + \log(1-\gamma) \Big) - \lambda D_n(h, g),$$

thus, model (4) is a penalized nonparametric model which penalizes h over departures from the known g, with tuning parameter λ.

2.2 Model estimation. As $h(\cdot)$ is unknown nuisance function, a common method to eliminate the infinite dimensional nuisance parameter h is to find the profile log-likelihood

$$\tilde{\ell}_n(\boldsymbol{\theta}, \Delta) = \sup_h \ell_n(\boldsymbol{\theta}, \Delta, h).$$

It is known that such supremum is a step function with jumps at the observed data points, i.e., the mass of $h(\cdot)$ has to be concentrated on the observations. So we will maximize the log-likelihood over $h(\cdot)$ that are step functions at observed values. The Cox proportional hazards model is a typical example of such, in which the log-likelihood function is maximized over the nonparametric base line step hazard functions, to get the 'partial likelihood' which is a profile likelihood. Thus we set

$h(y_i - \boldsymbol{\beta}'\mathbf{x}_i) = h_{i,0}, \quad h(y_i - \boldsymbol{\beta}'\mathbf{x}_i - \eta) = h_{i,1}, \quad h_i = \delta_i h_{i,1} + (1-\delta_i)h_{i,0}, \quad (i = 1, ..., n);$

then the corresponding 'complete' data log-likelihood is

$$\ell_n(\boldsymbol{\theta}, \Delta, \mathbf{h}) = \sum_{i=1}^{n} \Big\{ \delta_i \Big((1-\lambda) \log h_{i,1} + \lambda \log g(y_i - \boldsymbol{\beta}'\mathbf{x}_i - \eta) + \log \gamma \Big)$$

$$+ (1-\delta_i) \Big((1-\lambda) \log h_{i,0} + \lambda \log g(y_i - \boldsymbol{\beta}'\mathbf{x}_i) + \log(1-\gamma) \Big) \Big\} - n \log \Delta. \qquad (5)$$

To eliminate the nuisance parameters $(h_{i,0}, h_{i,1})$'s, for fixed $\boldsymbol{\theta}$ and Δ, we maximize the above log-likelihood over $\mathbf{h} = (h_{1,0}, h_{1,1}, ..., h_{n,0}, h_{n,1})$ subject to

$$\sum_{i=1}^{n} h_i = 1, \quad \text{and} \quad \sum_{i=1}^{n} h_i^{1-\lambda} g^{\lambda}(y_i - \boldsymbol{\beta}'\mathbf{x}_i - \delta_i \eta) = \Delta \qquad (6)$$

and use the Lagrange multipliers, for fixed $\boldsymbol{\theta}$ and Δ we maximize

$$\ell_n(\boldsymbol{\theta}, \Delta, \mathbf{h}) + \zeta\big(1 - \sum_{i=1}^{n} h_i\big) - n\eta\big(\sum_{i=1}^{n} h_i^{1-\lambda} g^{\lambda}(y_i - \boldsymbol{\beta}'\mathbf{x}_i - \delta_i \eta) - \Delta\big) \qquad (7)$$

over $\mathbf{h}$, and get its estimate as solution $\hat{\mathbf{h}} = (\hat{h}_{1,0}, \hat{h}_{1,1}, ..., \hat{h}_{n,0}, \hat{h}_{n,1})$ to the following equations (Appendix), with $C = \sum_{i=1}^{n} h_i^{1-\lambda}$,

$$h_{i,1} = \frac{1}{n} \frac{1}{1 + \eta h_i^{-\lambda} [g^{\lambda}(y_i - \boldsymbol{\beta}'\mathbf{x}_i - \eta) - C^{-1}\Delta]}$$

$$\approx \frac{1}{n} \frac{1}{1 + \eta_0 h_i^{-\lambda} [g^{\lambda}(y_i - \boldsymbol{\beta}'\mathbf{x}_i - \eta) - C^{-1}\Delta]}, \qquad (8)$$

$$h_{i,0} = \frac{1}{n} \frac{1}{1 + \eta h_i^{-\lambda} [g^{\lambda}(y_i - \boldsymbol{\beta}'\mathbf{x}_i) - C^{-1}\Delta]} \approx \frac{1}{n} \frac{1}{1 + \eta_0 h_i^{-\lambda} [g^{\lambda}(y_i - \boldsymbol{\beta}'\mathbf{x}_i) - C^{-1}\Delta]}.$$

The $(h_{i,0}, h_{i,1})$'s can be solved iteratively.

Plugging in $(\hat{h}_{i,0}, \hat{h}_{i,1})$ into (4), we get the profile log-likelihood

$$\tilde{\ell}_n(\boldsymbol{\theta}, \Delta, \hat{\mathbf{h}}) = \sum_{i=1}^{n} \Big\{ \delta_i \Big((1-\lambda) \log \hat{h}_{i,1} + \lambda \log g(y_i - \boldsymbol{\beta}'\mathbf{x}_i - \eta) + \log \gamma \Big)$$

$$+ (1-\delta_i) \Big((1-\lambda) \log \hat{h}_{i,0} + \lambda \log g(y_i - \boldsymbol{\beta}'\mathbf{x}_i) + \log(1-\gamma)) \Big\} - n \log \Delta. \qquad (9)$$

For fixed Δ and $\hat{\mathbf{h}}$, the profile maximum likelihood estimate (MLE)[21,22] of $\boldsymbol{\theta}$ based on (9) is, with $\hat{\theta} = (\hat{\boldsymbol{\beta}}', \hat{\eta}', \hat{\lambda}, \hat{\gamma})'$,

$$\hat{\boldsymbol{\theta}} = \arg\sup_{\boldsymbol{\theta}} \tilde{\ell}_n(\boldsymbol{\theta}, \Delta, \hat{\mathbf{h}}).$$

However, since $\boldsymbol{\delta} = (\delta_1, ..., \delta_n)$ is latent (missing), we use the EM-algorithm[23], more specifically, the latent structure in Tan, Tian and Ng[24], to estimate $\hat{\mathbf{h}}$ and $\hat{\boldsymbol{\theta}}$. Let $\mathbf{Y}_n = (y_1, ..., y_n)$, $\mathbf{X}_n = (\mathbf{x}_1, ..., \mathbf{x}_n)$ and $\mathbf{h}^{(0)}$ and $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\beta}^{(0)}, \eta^{(0)}, \gamma^{(0)}, \lambda^{(0)})'$ be starting values of $\mathbf{h}$ and $\boldsymbol{\theta}$. Typically $h_{i,0}^{(0)} = h_{i,1}^{(0)} = 1/n$, $(\boldsymbol{\beta}^{(0)}, \eta^{(0)}, \gamma^{(0)})$ can be set as the MLE of $(\boldsymbol{\beta}, \eta, \gamma)$ under model $g(\cdot)$, and $\gamma^{(0)} = \lambda^{(0)} = 1/2$. At the r-th iteration, define in the E-step,

$$H_n(\mathbf{h}, \boldsymbol{\theta}, \Delta | \mathbf{h}^{(r)}, \boldsymbol{\theta}^{(r)}) = E_{\boldsymbol{\delta}}[\tilde{\ell}_n(\boldsymbol{\theta}, \Delta) | \mathbf{Y}_n, \mathbf{X}_n, \mathbf{h}^{(r)}, \boldsymbol{\theta}^{(r)}], \tag{10}$$

where the expectation is with respect to $\boldsymbol{\delta}$, and as if the true data is generated from parameters $\boldsymbol{\theta}^{(r)}$ and $\mathbf{h}^{(r)}$. The computation of (10) is given in the Appendix. In particular, the r-th step estimate of the δ_i's (for $i = 1, ..., n; r = 0, 1, 2, ...$), are

$$\delta_i^{(r)} = \frac{\gamma^{(r)} \big(h_{i,1}^{(r)}\big)^{1-\lambda^{(r)}} g^{\lambda^{(r)}}(y_i - \boldsymbol{\beta}'^{(r)} \mathbf{x}_i - \eta^{(r)})}{\gamma^{(r)} \big(h_{i,1}^{(r)}\big)^{1-\lambda^{(r)}} g^{\lambda^{(r)}}(y_i - \boldsymbol{\beta}'^{(r)} \mathbf{x}_i - \eta^{(r)}) + (1-\gamma^{(r)}) \big(h_{i,0}^{(r)}\big)^{1-\lambda^{(r)}} g^{\lambda^{(r)}}(y_i - \boldsymbol{\beta}'^{(r)} \mathbf{x}_i)},$$

and with $h_i^{(r)} = h_i^{(r)}(\boldsymbol{\theta}^{(r)}) = \delta_i^{(r)} h_{i,1}^{(r)}(\boldsymbol{\theta}^{(r)}) + (1 - \delta_i^{(r)}) h_{i,0}^{(r)}(\boldsymbol{\theta}^{(r)})$, set

$$\Delta^{(r)} = \Delta^{(r)}(\boldsymbol{\theta}^{(r)}) = \sum_{i=1}^{n} \big(h_i^{(r)}(\boldsymbol{\theta}^{(r)})\big)^{1-\lambda^{(r)}} g^{\lambda^{(r)}}\big(y_i - \boldsymbol{\beta}'^{(r)} \mathbf{x}_i - \delta_i^{(r)} \eta^{(r)}\big), \quad (r = 0, 1, 2, ...).$$

In the M-step, for fixed $\boldsymbol{\theta}^{(r)}$ and $\Delta^{(r)}$, define

$$\mathbf{h}^{(r+1)} = \mathbf{h}^{(r+1)}(\boldsymbol{\theta}) = \arg\sup_{\mathbf{h}} H_n(\mathbf{h}, \boldsymbol{\theta}, \Delta^{(r)} | \mathbf{h}^{(r)}, \boldsymbol{\theta}^{(r)}) \quad \text{subject to (6)},$$

which is just the maximization of (4) with the δ_i's replaced by the $\delta_i^{(r)}$'s, over $\mathbf{h}$ subject to (6) and with Δ replaced by $\Delta^{(r)}$. In particular, for fixed $\boldsymbol{\theta}$ and $\Delta^{(r)}$, $\mathbf{h}^{(r+1)} = \mathbf{h}^{(r+1)}(\boldsymbol{\theta})$ can be obtained using (8). Let

$$\boldsymbol{\theta}^{(r+1)} = \arg\sup_{\boldsymbol{\theta}} H_n(\mathbf{h}^{(r+1)}, \boldsymbol{\theta}, \tilde{\Delta} | \mathbf{h}^{(r)}, \boldsymbol{\theta}^{(r)}).$$

The above optimization is the same as maximizing (4) over $\boldsymbol{\theta}$, in which $(h_{i,0}, h_{i,1})$'s are replaced by $(h_{i,0}^{(r+1)}, h_{i,1}^{(r+1)})$'s, and with $\Delta = \Delta(\boldsymbol{\theta})$ approximated by

$$\tilde{\Delta}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left(h_i^{(r)}(\boldsymbol{\theta}^{(r)})\right)^{1-\lambda^{(r)}} g^{\lambda^r}(y_i - \boldsymbol{\beta}'\mathbf{x}_i - \delta_i^{(r)}\eta).$$

Then it is known (e.g., Dempster et al.[23], 1997) that as $r \to \infty$ one has

$$\boldsymbol{\theta}^{(r)} \to \hat{\boldsymbol{\theta}}, \qquad \Delta^{(r)} \to \hat{\Delta}.$$

2.3 Asymptotic properties. The classification is consistent only if the corresponding parameter estimation is consistent. To study the asymptotic behavior of the estimators, the following notations will be used. Below we study the almost sure consistency of our profile MLE $\hat{\boldsymbol{\theta}}$ under relatively simply conditions. Let

$$\mathbf{i}_{0,1} = -\Big(\gamma\lambda g^{\lambda-1}(y - \boldsymbol{\beta}'\mathbf{x} - \eta)h^{1-\lambda}(y - \boldsymbol{\beta}'\mathbf{x} - \eta)\dot{g}(y - \boldsymbol{\beta}'\mathbf{x} - \eta)$$

$$+\gamma(1-\lambda)g^{\lambda}(y - \boldsymbol{\beta}'\mathbf{x} - \eta)h^{-\lambda}(y - \boldsymbol{\beta}'\mathbf{x} - \eta)\dot{h}(y - \boldsymbol{\beta}'\mathbf{x} - \eta)\Big)/f(\epsilon|\boldsymbol{\theta}, h),$$

$$\mathbf{i}_{0,2} = -\Big((1-\gamma)\lambda g^{\lambda-1}(y - \boldsymbol{\beta}'\mathbf{x})h^{1-\lambda}(y - \boldsymbol{\beta}'\mathbf{x})\dot{g}(y - \boldsymbol{\beta}'\mathbf{x})$$

$$+(1-\gamma)(1-\lambda)g^{\lambda}(y - \boldsymbol{\beta}'\mathbf{x})h^{-\lambda}(y - \boldsymbol{\beta}'\mathbf{x})\dot{h}(y - \boldsymbol{\beta}'\mathbf{x})\Big)/f(\epsilon|\boldsymbol{\theta}, h),$$

and $\mathbf{z}_1 = (\mathbf{x}', 1)'$, $\mathbf{z}_2 = (\mathbf{x}', 0)'$. Let $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0, \eta_0, \lambda_0)$ be the true parameter values for the observed data.

We need the following conditions

(C1). For fixed Δ, the log-likelihood (4) has a unique profile MLE $\hat{\boldsymbol{\theta}}$.

(C2). $\int \sqrt{f_0(\epsilon)}d\epsilon < \infty$.

(C3). $\{f(\cdot|\boldsymbol{\theta}, h) : \boldsymbol{\theta} \in \boldsymbol{\Theta}, h \in \mathcal{H}\}$ is bounded and continuous.

(C4). $h_0(\cdot)$ is uniformly continuous.

(C5). $\dot{g}(\cdot)$ and $\dot{h}(\cdot)$ are bounded continuous.

(C6). The $\Omega(\boldsymbol{\theta}_0)$ given in Theorem 2 below is non-singular.

Theorem 1. *Assume (C1)-(C4), then for the profile MLE* $\hat{\boldsymbol{\theta}}$*,*

$$\hat{\boldsymbol{\theta}} \stackrel{a.s.}{\to} \boldsymbol{\theta}_0.$$

Theorem 2. *Assume (C1),(C5) and (C6), then the profile MLE* $\hat{\boldsymbol{\theta}}$ *is efficient for* $\boldsymbol{\theta}$ *in the semiparametric model (2), and*

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0') \stackrel{D}{\to} N(\mathbf{0}, \Omega^{-1}(\boldsymbol{\theta}_0)), \qquad \Omega(\boldsymbol{\theta}_0) = E_{\boldsymbol{\theta}_0, h_0}[\mathbf{i}^* \mathbf{i}^{*'}],$$

where $\mathbf{i}^*$ *is the efficient score for* $\boldsymbol{\theta}$ *in the presence of the nuisance* $h(\cdot)$*. In particular,*

$$\sqrt{n}[(\hat{\boldsymbol{\beta}}', \hat{\eta}')' - (\boldsymbol{\beta}_0', \eta_0')'] \stackrel{D}{\to} N(\mathbf{0}, \Omega_0^{-1}(\boldsymbol{\theta}_0)), \qquad \Omega_0(\boldsymbol{\theta}_0) = E_{\boldsymbol{\theta}_0, h_0}[\mathbf{i}_0^* \mathbf{i}_0^{*'}],$$

where $\mathbf{i}_0^*$ *is the efficient score for* $(\boldsymbol{\beta}', \eta')'$ *in the presence of the nuisance* $h(\cdot)$ *given by* $\mathbf{i}_0^* = \mathbf{i}_{0,1}(\epsilon)(\mathbf{z}_1 - \boldsymbol{\mu}_1) + \mathbf{i}_{0,2}(\epsilon)(\mathbf{z}_2 - \boldsymbol{\mu}_2)$*,* $\boldsymbol{\mu}_1 = E_{\boldsymbol{\theta}_0}(\mathbf{z}_1)$ *and* $\boldsymbol{\mu}_2 = E_{\boldsymbol{\theta}_0}(\mathbf{z}_2)$*.*

2.4 Testing the null hypothesis. To test $H_0 : \eta = 0$ vs the alternative $H_1 : \eta \neq 0$ under parametric model, we commonly use test statistics including the likelihood ratio statistic, score test statistic and the Wald statistic. Generally the former two statistics are preferred for invariance of parameterization and one approximation (the limit chi-squared distribution), while the Wald statistic has two approximations (the limit chi-squared distribution, and the asymptotic variance of the estimator). However, under H_0, γ in model (2) is non-identifiable, simply omitting γ will result in a model which is not a sub-model of model (2), hence the classical likelihood ratio

test and the score cannot be applied here. So we use the Wald statistic which requires the estimator to be computed under the alternative and dose not have model identifiability problem.

Generally, denote $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ with $dim(\boldsymbol{\theta}) = d$ and $dim(\boldsymbol{\theta}_1) = d_1$, and $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)$ be the MLE of $\boldsymbol{\theta}$ under the full model. Consider the null hypothesis $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_{1,0}$. The Wald test statistic is given by

$$W_n = (\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{1,0})' Cov^{-1}(\hat{\boldsymbol{\theta}}_1)(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{1,0}).$$

If $Cov(\hat{\boldsymbol{\theta}}_1)$ is known, $W_n \sim \chi^2_{d_1}$. If $Cov(\hat{\boldsymbol{\theta}}_1)$ is estimated, $W_n/d_1 \sim F_{d_1,n-d}$.

In our problem, $\boldsymbol{\theta}_1 = \eta$, $\eta_{1,0} = 0$, treat $Cov(\hat{\eta}_n)$ as known, so $W_n = \hat{\eta}_n Cov(\hat{\eta}_n)\hat{\eta}_n \sim \chi^2_1$ asymptotically, and if $W_n > \chi^2_1(1-\alpha)$, the upper $(1-\alpha)$-th quantile of the χ^2_1 distribution, then H_0 is rejected.

2.5 Classification rule. After the null hypothesis above is rejected, we classify subject i into one of the groups. We first evaluate $h(y_i - \hat{\boldsymbol{\beta}}' \mathbf{x}_i - \hat{\eta})$ and $h(y_i - \hat{\boldsymbol{\beta}}' \mathbf{x}_i)$. If subject i is one of those used in the computation, we have $\hat{h}_{i,1}$ and $\hat{h}_{i,0}$ for the mentioned two values. But if subject i is a new individual, we do not have these values. Thus we first interpolate the values of $\{\hat{h}_{i,1}, \hat{h}_{i,0} : i = 1, ..., n\}$ on the points $\{y_i - \hat{\boldsymbol{\beta}}' \mathbf{x}_i - \hat{\eta}, y_i - \hat{\boldsymbol{\beta}}' \mathbf{x}_i : i = 1, ..., n\}$, to the function $h_n(\cdot)$ on R. Note that for 1-dimensional response, $h_n(\cdot)$ is just the linear interpolation (or polynomial, spline interpolation). For two or three dimensions, bilinear and trilinear interpolations are used. Note, though, that these interpolants are no longer linear functions of the spatial coordinates, rather products of linear functions; this are illustrated by the clearly non-linear example of bilinear interpolation in the figure below. Other extensions of linear interpolation can be applied to other kinds of mesh such as triangular and tetrahedral meshes, including the Bézier surfaces. These may be defined as

indeed higher-dimensional piecewise linear function. Since $h_n(\cdot)$ is interpolation of $\{\hat{h}_{i,1}, \hat{h}_{i,0} : i = 1, ..., n\}$ on $\{y_i - \hat{\boldsymbol{\beta}}'\mathbf{x}_i - \hat{\eta}, y_i - \hat{\boldsymbol{\beta}}'\mathbf{x}_i : i = 1, ..., n\}$,

$$h_n(y_i - \hat{\boldsymbol{\beta}}'\mathbf{x}_i - \hat{\eta}_i) = \hat{h}_{i,1}, \quad h_n(y_i - \hat{\boldsymbol{\beta}}'\mathbf{x}_i) = \hat{h}_{i,0}, \quad (i = 1, ..., n).$$

Note that $h_n(\cdot)$ may not be a proper density function, but this does not matter, as we will see that only the ratios of $h_n(y_i - \hat{\boldsymbol{\beta}}'\mathbf{x}_i - \hat{\eta}_i)/h_n(y_i - \hat{\boldsymbol{\beta}}'\mathbf{x}_i)$'s will be used in the classification. However, if the additive mixture model is used instead of the geometric mixture (2), then we need to normalize $h_n(\cdot)$ to make it a proper density function in the classification procedure below, and such normalization can be non-trivial for multi-dimensional function.

For each subject $i = 1, 2, ..., n$ with treatment j, this subject can be classified to subgroup S_{j1} ($\delta_i = 1$) by the commonly used classical Bayesian classification rule, that is, if $P(\delta_i = 1|y_i, \mathbf{x}_i, \hat{\mathbf{h}}, \hat{\boldsymbol{\theta}}) > 1/2$ or

$$\hat{\gamma}[g^{\hat{\lambda}}h_n^{1-\hat{\lambda}}](y_i - \hat{\boldsymbol{\beta}}'\mathbf{x}_i - \hat{\eta}) > (1 - \hat{\gamma})[g^{\hat{\lambda}}h_n^{1-\hat{\lambda}}](y_i - \hat{\boldsymbol{\beta}}'\mathbf{x}_i),$$

otherwise classify this subject to subgroup S_{j0} ($\delta_i = 0$). Therefore, we obtain a collection $\{S_{jr} : j = 1, ..., k; r = 0, 1\}$ of subgroups, where each S_{j1} is the favorable subgroup of the j-th treatment; and S_{j0} is the non-favorable subgroup.

However, as mentioned in the Introduction, the correct classification of the treatment favorable subgroup is of more clinical importance, and the Bayesian rule does not take this into consideration. So we use the Neyman-Pearson classification rule, such that the mis-classification error rate of the treatment favorable subgroup is under control at pre-specified level α, while the misclassification error for the other subgroup is minimized. To be specific, for each subject i, denote the likelihood ratio

$$LR(y_i, \mathbf{x}_i) = \frac{f(y_i, \mathbf{x}_i|\hat{\boldsymbol{\theta}}, \delta = 1)}{f(y_i, \mathbf{x}_i|\hat{\boldsymbol{\theta}}, \delta = 0)} = \frac{[g^{\hat{\lambda}}h_n^{1-\hat{\lambda}}](y_i - \hat{\boldsymbol{\beta}}'\mathbf{x}_i - \hat{\eta})}{[g^{\hat{\lambda}}h_n^{1-\hat{\lambda}}](y_i - \hat{\boldsymbol{\beta}}'\mathbf{x}_i)}.$$

Parallel to the NP uniformly most powerful test procedure for testing the simple hypothesis $H_0 : \eta = 0$ vs $H_1 : \eta \neq 0$, the optimal classification rule is: classify the i-th subject to subgroup S_1 if

$$LR(y_i, \mathbf{x}_i) > k(\alpha), \quad \text{with } k(\alpha) \text{ determined by } \quad P_{H_0}\big(LR(y, \mathbf{X}) \geq k(\alpha)\big) = \alpha.$$

To find an approximate solution for $K(\alpha)$, let $\{LR_j : j = 1, ..., n_0\}$ be the LR_s of patients who are in the unfavorable subgroup (for simulated data, the subgroup memberships are known), then set $K(\alpha)$ to be the $(1-\alpha)$-th upper quantile of $LR_1, ..., LR_{n_0}$. Note generally there is no monotonic relationship between α and $K(\alpha)$. In fact, value of $K(\alpha)$ depends on α, $\boldsymbol{\beta}_0$, λ_0, γ_0 and the shape of $h(\cdot)$. These quantities change from data to data. However, for real data $\{(y_i, \mathbf{x}_i) : i = 1, ..., n\}$, the subgroup memberships are unknown, we cannot use the above method to decide $K(\alpha)$. To obtain $K(\alpha)$, set $LR_i = \hat{f}_n(\epsilon_i - \hat{\alpha})/\hat{f}_n(\epsilon_i)$, and let

$$Q_n(t) = \sum_{i=1}^{n} w_{ni} I(LR_i \leq t), \quad w_{ni} = (1 - \hat{\delta}_i)/\sum_{j=1}^{n}(1 - \hat{\delta}_j)$$

be the weighted empirical distribution of the LR_i's, then set $K(\alpha) = Q_n^{-1}(1-\alpha)$ to be the $(1-\alpha)$-th upper quantile of Q_n. For new patient with covariate $\mathbf{x}$ but without response y, we define

$$LR(\mathbf{x}) = E_{H_0}\Big(\frac{f(y - \hat{\boldsymbol{\beta}}'\mathbf{x} - \hat{\alpha})}{f(y - \hat{\boldsymbol{\beta}}'\mathbf{x})}\Big|\mathbf{x}\Big) \approx \frac{1}{n_0}\sum_{i=1}^{n_0} \frac{f(y_i - \hat{\boldsymbol{\beta}}'\mathbf{x} - \hat{\alpha})}{f(y_i - \hat{\boldsymbol{\beta}}'\mathbf{x})}$$

and classify this patient to group 1 if $LR(\mathbf{x}) > K(\alpha)$, with $K(\alpha)$ given above.

We remark that the miss-classification errors are in the sense of conditioning on existence of the subgroups, i.e., the rejection of the null hypothesis. Such conclusion has a type II error, and so the nominal clasification error should be evaluated in the same way as the two-stage clinical trail, such as in Gao, Roy and Tan [30].

3. Simulation study and ral world data analysis

3.1 Simulation study.

We simulate $n = 1000$ i.i.d. samples with 1-dimensional response y_i's and with covariates $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3})$. We first generate the covariates, sample the $\mathbf{x}_i$'s from the 3-dimensional normal distribution with mean vector $\boldsymbol{\mu} = (3.1, 1.8, -0.5)'$ and covariance matrix Γ, with

$$\Gamma^{1/2} = \begin{pmatrix} 0.73 & -0.07 & 0.55 \\ 1.34 & -0.14 & 0.57 \\ 1.52 & -0.37 & 1.53 \end{pmatrix}.$$

Then we generate the response data, which, given the covariates, are from the mixture $\Delta_0^{-1} g^{\lambda_0} h^{1-\lambda_0}$. The y_i's are generated as

$$y_i = \boldsymbol{\beta}_0 \mathbf{x}_i + \delta_i \eta_0 + \epsilon_i, \quad (i = 1, ..., n).$$

We will show estimation results with six different choices of $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0, \eta_0, \gamma_0, \lambda_0)$ and compare the corresponding results from the commonly used normal mixture model, with $\phi(\cdot|\sigma^2)$ be the density of $N(0, \sigma^2)$,

$$f(\epsilon) = \gamma\phi(y - \boldsymbol{\beta}'\mathbf{x} - \eta|\sigma^2) + (1 - \gamma)\phi(y - \boldsymbol{\beta}'\mathbf{x}|\sigma^2).$$

Let g as the density of $N(0, \sigma^2)$. In the simulation, h is the density of Gamma(k, θ) distribution $(k, \theta) = (2, 1)$, resulting a mixture model with a skewed distribution, and deviating from the assumption of a normal model.

The geometric mixture model (2) has an innate relationship with the additive mixture model

$$f(\epsilon|\delta) = \gamma g(\epsilon|\delta) + (1 - \gamma)h_0(\epsilon|\delta)$$

as described by the following

Proposition. *For any given* $(\rho, g(\cdot), h_0(\cdot))$ *with* $(0 \le \rho \le 1)$*, there are* λ *and* $h(\cdot)$ *(a density function) such that*

$$\rho g(\cdot) + (1-\rho)h_0(\cdot) \equiv C(\lambda) g^{\lambda}(\cdot) h^{1-\lambda}(\cdot), \qquad C(\lambda) = \int g^{\lambda}(\epsilon) h^{1-\lambda}(\epsilon) d\epsilon.$$

Also, the correspondence between η *and* λ *is monotonic, and* $\rho = 0, 1$ *iff* $\lambda = 0, 1$.

We use the Proposition to sample the ϵ_i's, ie. sample them from the additive mixture model

$$\lambda_0 g(\epsilon) + (1-\lambda_0) h(\epsilon).$$

The above sampling is straight forward. When ϵ_i is from $h(\epsilon)$, replace ϵ_i by $\epsilon_i - E(\epsilon)$, as the ϵ_i's needed to be zero mean valued. Although the λ_0 from this model is not the same as from model (3), it corresponds to some λ_0 from model (3), we can find their correspondence for some pairs of values and estimate it from model (3).

Then with the simulated data $(y_i, \mathbf{x}_i)$'s, we first fit model (3) – actually model (4) treating the δ_i's as missing data via the EM-algorithm. In particular, we set the starting values as $\mathbf{h}^{(0)} = (h_1^{(0)}, ..., h_n^{(0)}) = (1/n, ..., 1/n)$, $(\boldsymbol{\beta}^{(0)}, \eta^{(0)}, \lambda^{(0)}) = (\boldsymbol{\beta}_0/2, \eta_0/2, 0.5)$. Then compute $(\mathbf{h}^{(r)}, \boldsymbol{\theta}^{(r)}, \delta_i^{(r)})$'s by the EM-algorithm. Convergence of the algorithm can be accessed by the criterion, with a given ρ (typically $\rho = 10^{-3}$),

$$\frac{||\boldsymbol{\theta}^{(r+1)} - \boldsymbol{\theta}^{(r)}||}{||\boldsymbol{\theta}^{(r)}||} \le \rho.$$

When the above criterion is met at the $(r+1)$-th iteration, the EM algorithm is stopped, and $\boldsymbol{\theta}^{(r+1)}$ is treated as the profile MLE $\hat{\boldsymbol{\theta}}$.

Below, in Table 1 (A and B), for six different combinations of $\boldsymbol{\theta}_0$ values, we show the estimation results based on the normal model and the proposed model (mixture), *sd* is the estimated standard deviation of the estimator, which is computed by 1000 repetitions, as this is much easier than to compute it from the Hessian of the profile likelihood. Table 1A gives the results using the true parameter values and estimates

with estimated standard errors, while Table 1B displays the same results using biases and mean squared errors, based on the same data sets. As expected the profile model gives parameter estimates closer to the true values and smaller biases than does the normal model.

For the normal model, there is no λ, so the corresponding column is blank. MLE (profile) is the MLE of $\boldsymbol{\theta}_0$ under the proposed profile likelihood, MLE(normal) is the MLE of $\boldsymbol{\theta}_0$ under the normal model; $[sd]$ is the estimated standard deviation of the estimated parameters. From the above table, when the true model deviated from the normal model, for most components of $\boldsymbol{\theta}_0$, the estimates from the semirarametric mixture model are significantly better than those from the normal model. However, the standard errors from both the profile likelihood model and the normal model, vary, i.e., none of the model is uniformly better than the other in this respect, likely because the profile model is more flexible and is semiparametric, while the normal model is parametric. It is known that parameter estimate from a semiparametric model generally has bigger variance than that from a parametric model. But when the parametric model is far from the true model, the variance of its parameter estimate can be bigger than that of a semiparametric model. Such results vary from data sets to data sets, and even within the same data set for different parameter components.

In Table 2, we give the results of testing whether subgroups even exist ($H_0 : \eta = 0$ vs $H_1 : \eta \neq 0$) using the proposed semiparametric model for the 3 datasets with small η values and the same 6 datasets used in Table 1. In addition, Table 2 gives the N-P error (the proportion of mis-classified subjects in group 1: $P(\text{Reject } H_0)$), the overall error (the proportion of all mis-classified subjects for two subgroups), and the cut-off point $K(0.05)$ corresponding to level 0.05. The simulation has 1000 repetitions. From Table 2, the model is able to detect even very small subgroup effect, e.g., $\eta_0 = 0.01, 0.02$, and to keep the misclassification error for group 1 under

Table 1A. Parameter estimates under two models (simulated data)

$\boldsymbol{\theta}$	β	η	σ	γ	λ
$\boldsymbol{\theta}_0$	(-1.5, 2.5, 1.0)	0.85	2.1	0.6	0.6
MLE(profile)	(-1.314, 2.356, 1.016)	0.730	1.972	0.726	0.622
[sd]	[0.009, 0.016, 0.036]	[0.023]	[0.078]	[0.062]	[0.014]
MLE(normal)	(-1.506, 2.116, 1.241)	4.572	2.448	0.539	
[sd]	[0.008, 0.110, 0.094]	[0.197]	[0.073]	[0.024]	
$\boldsymbol{\theta}_0$	(1.3, 1.2, -1.6)	1.65	2.1	0.5	0.7
MLE(profile)	(1.374, 1.161, -1.642)	2.164	2.140	0.463	0.559
[sd]	[0.077, 0.105, 0.034]	[0.059]	[0.049]	[0.023]	[0.009]
MLE(normal)	(1.287, 0.852, -1.340)	4.583	1.941	0.523	
[sd]	[0.009, 0.050, 0.070]	[0.104]	[0.059]	[0.015]	
$\boldsymbol{\theta}_0$	(-1.5, 2.5, 1.0)	3.5	2.1	0.6	0.6
MLE(profile)	(-1.237, 2.483, 0.960)	3.967	2.154	0.468	0.453
[sd]	[0.083, 0.159, 0.104]	[0.141]	[0.046]	[0.032]	[0.010]
MLE(normal)	(-1.496, 2.495, 0.987)	5.330	2.250	0.510	
[sd]	[0.009, 0.070, 0.101]	[0.112]	[0.074]	[0.011]	
$\boldsymbol{\theta}_0$	(-1.5, 2.5, 1.0)	6.35	2.1	0.6	0.6
MLE(profile)	(-1.469, 2.486, 0.995)	6.637	2.152	0.617	0.581
[sd]	[0.067, 0.081, 0.050]	[0.155]	[0.054]	[0.027]	[0.009]
MLE(normal)	(-1.493, 2.604, 0.932)	6.855	2.414	0.530	
[sd]	[0.005, 0.056, 0.089]	[0.120]	[0.085]	[0.013]	
$\boldsymbol{\theta}_0$	(1.2, -1.4, 3.2)	7.74	2.1	0.4	0.3
MLE(profile)	(1.368, -1.403, 3.124)	7.820	2.487	0.420	0.358
[sd]	[0.131, 0.173, 0.075]	[0.155]	[0.037]	[0.015]	[0.006]
MLE(normal)	(1.206, -1.225, 3.080)	8.202	2.480	0.550	
[sd]	[0.008, 0.049, 0.060]	[0.207]	[0.057]	[0.017]	
$\boldsymbol{\theta}_0$	(-2.3, -1.4, 2.1)	8.39	2.1	0.7	0.8
MLE(profile)	(-2.271, -1.429, 2.096)	8.523	2.172	0.700	0.777
[sd]	[0.060, 0.093, 0.041]	[0.068]	[0.067]	[0.016]	[0.008]
MLE(normal)	(-2.295, -1.352, 2.062)	8.618	2.287	0.514	
[sd]	[0.003, 0.042, 0.047]	[0.125]	[0.100]	[0.010]	

Table 1B. Parameter estimates and biases under two models (simulated data)

$\boldsymbol{\theta}$	β	η	σ	γ	λ
$\boldsymbol{\theta}_0$	(-1.5, 2.5, 1.0)	0.85	2.1	0.6	0.6
bias(profile)	(0.186,-0.144,0.016)	-0.120	-0.128	0.126	0.022
MSE	[0.035, 0.021, 0.002]	[0.015]	[0.022]	[0.020]	[0.001]
bias(normal)	(-0.006, -0.384, 0.241)	3.722	0.348	-0.061	
MSE	[0.000, 0.160, 0.067]	[13.892]	[0.126]	[0.004]	
$\boldsymbol{\theta}_0$	(1.3, 1.2, -1.6)	1.65	2.1	0.5	0.7
bias(profile)	(0.074, -0.039,-0.042)	0.514	0.040	-0.037	-0.141
MSE	[0.011,0.013,0.003]	[0.268]	[0.004]	[0.002]	[0.020]
bias(normal)	(-0.013,-0.348,0.260)	2.933	-0.159	0.023	
MSE	[0.000, 0.124, 0.073]	[8.613]	[0.029]	[0.001]	
$\boldsymbol{\theta}_0$	(-1.5, 2.5, 1.0)	3.5	2.1	0.6	0.6
bias(profile)	(0.263,-0.017,-0.040)	0.467	0.054	-0.132	-0.147
MSE	[0.076,0.026,0.012]	[0.238]	[0.005]	[0.018]	[0.022]
bias(normal)	(0.004,-0.005,-0.013)	1.830	0.150	-0.090	
MSE	[0.000, 0.005, 0.010]	[3.361]	[0.028]	[0.008]	
$\boldsymbol{\theta}_0$	(-1.5, 2.5, 1.0)	6.35	2.1	0.6	0.6
bias(profile)	(0.031,-0.014,-0.005)	0.287	0.052	0.017	-0.019
MSE	[0.005, 0.007, 0.003]	[0.106]	[0.006]	[0.001]	[0.000]
bias(normal)	(0.007,0.104,-0.068)	0.505	0.314	-0.070	
MSE	[0.000, 0.014, 0.013]	[0.269]	[0.106]	[0.005]	
$\boldsymbol{\theta}_0$	(1.2, -1.4, 3.2)	7.74	2.1	0.4	0.3
bias(profile)	(0.168, -0.003, -0.076)	0.080	0.387	0.020	0.058
MSE	[0.045,0.030,0.011]	[0.030]	[0.151]	[0.001]	[0.003]
bias(normal)	(0.006, 0.175, -0.120)	0.462	0.380	0.150	
MSE	[0.000, 0.033, 0.018]	[0.256]	[0.148]	[0.023]	
$\boldsymbol{\theta}_0$	(-2.3, -1.4, 2.1)	8.39	2.1	0.7	0.8
bias(profile)	(0.029,-0.029,-0.004)	0.133	0.072	0.000	-0.023
MSE	[0.004, 0.009, 0.002]	[0.022]	[0.010]	[0.000]	[0.001]
bias(normal)	(0.005,0.048,-0.038)	0.228	0.187	-0.186	
MSE	[0.000, 0.004, 0.004]	[0.068]	[0.045]	[0.035]	

Table 2. Hypothesis test and classification results (simulated data)

η_0	$\hat{\eta}$	P(Reject H_0)	Overall Error	N-P Error	$K(0.05)$
0.00	0.000	0.06			
0.01	0.009	0.98	0.500	0.052	1.026
0.02	0.017	1	0.500	0.052	1.053
0.85	0.730	1	0.582	0.054	3.636
1.65	2.164	1	0.426	0.052	29.177
3.50	3.967	1	0.442	0.054	5.587
6.35	6.637	1	0.168	0.054	32.420
7.74	7.820	1	0.050	0.052	1.489
8.39	8.523	1	0.028	0.055	0.269

the prespecified level (0.05). Also since different datasets are used in the table, the

K value could be quite different.

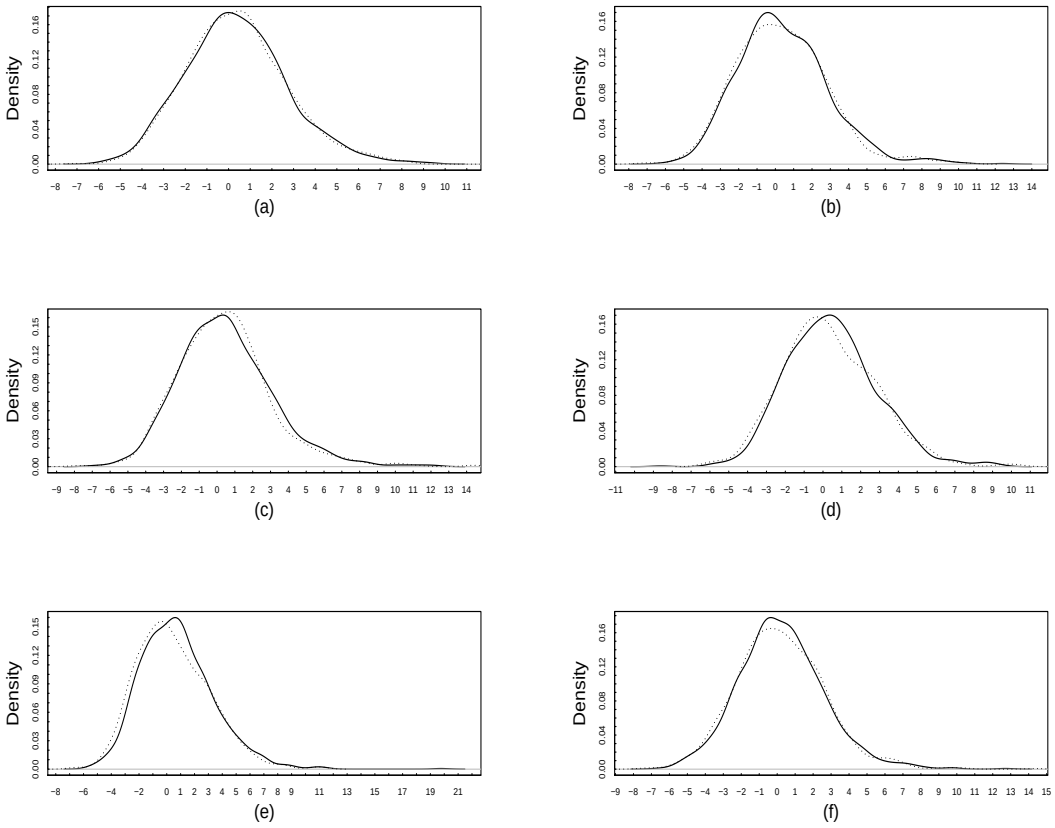

Figure 1: Densities for different models.

Figure 1. presents the true densities (solid lines) and estimated profile densities (dotted lines) of ϵ. Panel (a)-(f) shows the corresponding densities for the six parameter sets in Table 1.

3.2 Application to a real world trial.

ACTG175 is a randomized, double-blind trial coordinated by the AIDS Clinical Trials Group (ACTG), and supported by the National Institute of Allergy and

Infectious Diseases (NIAID). ACTG175 was conducted at 43 sites within both the adult and pediatric ACTG's and 9 sites of the National Hemophilia Foundation. Participants were enrolled into the study between December 1991 and October 1992, and received treatment through December 1994. Follow-up and final evaluations of participants took place between December 1994 and February 1995.

We analyze this data using the proposed method for each of the four treatments for subgroup analysis purpose. Treatment 1 is the ZDV therapy, treatment 2 is the combined therapy ZDV+ddI, treatment 3 is the combined therapy ZDV+zalcitabine, treatment 4 is ddI. The number of patients in the four treatments are 532,522,524 and 561, respectively. In all the four analyses, the response variable is the CD4 counts after 20 weeks of the corresponding treatment, and the covariates are Age, baseline CD4 counts and Gender, the corresponding coefficients are $(\beta_1, \beta_2, \beta_3)$. The estimated standard deviations are obtained via bootstrap (Table 3-4).

The analysis (Table 3) shows that the four treatments all have significant differential effects on some subgroup. Most of the parameter estimates are close for the proposed model and the mixture normal model, but the estimates on η differ a lot, which suggest that the mixture normal model is less accurately estimated and more variable. For the normal mixture model, the coefficients of variation on η are 70%, 16%, 38% and 20% for treatment 1-4 respectively, whereas for the proposed model, they are only 16%, 11%, 2%, and 12%. Our algorithm classified patients into different subgroups. For each treatment of the four treatments, the null hypothesis H_0 is rejected, implying there exists subgroup of patients who benefit more than others (Table 4). However, the estimated group 1 proportions (γ) is very small (about 5%) for each treatment, suggesting each one of the four treatments is especially efficacious for only a very small percentage of patients. This is consistent with the evolution HIV treatments ZDV, ZDV+ddI, ZDV+zalcitabine, and ddI, where all have worked

Table 3. Parameter estimates (real data)

(β_1,β_2 and β_3 are coefficients for Age, baseline CD4 counts and Gender)

$\boldsymbol{\theta}$	β_0	$(\beta_1, \beta_2, \beta_3)$	η	σ	γ	λ
Trtmnt1						
MLE(profile)	91.980	(-0.569, 0.748, -2.805)	18.194	0.754	0.011	0.745
[sd]	[30.922]	[0.603, 0.033, 2.777]	[2.984]	[0.023]	[0.003]	[0.020]
MLE(normal)	95.440	(-0.325,0.714,-9.880)	170.301	0.717	0.034	
[sd]	[19.819]	[0.512, 0.052, 9.987]	[120.61]	[0.019]	[0.020]	
Trtmnt2						
MLE(profile)	128.110	(1.774, 0.610, -3.025)	16.996	0.854	0.002	0.693
[sd]	[33.368]	[0.728, 0.058, 2.011]	[1.976]	[0.028]	[0.001]	[0.012]
MLE(normal)	157.252	(1.036,0.601,-10.202)	436.397	0.794	0.015	
[sd]	[28.056]	[0.496,0.067,2.085]	[69.133]	[0.029]	[0.005]	
Trtmnt3						
MLE(profile)	178.267	(-1.480, 0.698, -4.049)	119.633	0.785	0.008	0.686
[sd]	[33.177]	[0.764, 0.050, 2.517]	[2.028]	[0.029]	[0.005]	[0.092]
MLE(normal)	183.151	(-1.652,0.701,-3.574)	509.752	0.741	0.007	
[sd]	[12.823]	[0.431,0.032,3.494]	[196.888]	[0.019]	[0.006]	
Trtmnt4						
MLE(profile)	100.514	(-0.405, 0.829, -2.246)	17.255	0.764	0.008	0.74300
[sd]	[20.202]	[0.560, 0.020, 2.428]	[2.219]	[0.021]	[0.014]	[0.014]
MLE(normal)	111.332	(-0.600, 0.818, -1.862)	297.652	0.746	0.007	
[sd]	[26.762]	[0.449, 0.051, 1.042]	[61.071]	[0.028]	[0.003]	

Table 4. Hypothesis test results (real data)

	Decision	$K(0.05)$	group1-percent
Trtmnt1	H_1	5.591	0.053
Trtmnt2	H_1	8.792	0.050
Trtmnt3	H_1	4.901	0.046
Trtmnt4	H_1	6.315	0.052

to some extent for most of the patients. With the method, it is conceivable that we can further analyze the 5% patients identified to plan and improve future therapeutic development.

4. Discussion.

We have proposed a semiparametric model for the analysis of subgroups for clinical trial study. The specified model has a known parametric component taking into account covariates, and an unknown nonparametric component. The profile likelihood along with EM algorithm is utilized to estimate model parameters, including the subgroup effects. The null hypothesis of no subgroup is tested via the Wald statistic, and when the existence of subgroups is confirmed, the patients are classified to the subgroups by the Neyman-Pearson classification rule, which guarantees the misclassification error for the treatment favorable group under pre-specified controlled level. Simulation studies show the method outperform the parametric model when it is not correctly specified. The method is then used to analyze a real world clinical trial data, getting insights on the treatments. More broadly, we believe the proposed method provides a useful tool to advance precision medicine and to generate new hypotheses while at the same time providing some credence to subgroup analysis findings.

Appendix.

Derivation of (8). Denote $\sum_{i=1}^{n} h_i^{1-\lambda} = C$. Then the constraint $\sum_{i=1}^{n} h^{1-\lambda} g^{\lambda}(y_i - \boldsymbol{\beta}'\mathbf{x}_i - \delta_i\eta) = \Delta$ is re-written as $\sum_{i=1}^{n} h^{1-\lambda}[g^{\lambda}(y_i - \boldsymbol{\beta}'\mathbf{x}_i - \delta_i\eta) - C^{-1}\Delta] = 0$. Similarly as finding the optimal weights for empirical likelihood, Differentiating (7)

$$\ell_n(\boldsymbol{\theta}, \Delta, \mathbf{h}) + \zeta(1 - \sum_{i=1}^{n} h_i) - n\eta \sum_{i=1}^{n} h_i^{1-\lambda}[g^{\lambda}(y_i - \boldsymbol{\beta}'\mathbf{x}_i - \delta_i\eta) - C^{-1}\Delta]$$

over $h_{i,1}$ and $h_{i,0}$ and set them to zeros respectively, we get

$$\begin{cases} 0 = \frac{\delta_i(1-\lambda)}{h_{i,1}} + \zeta\delta_i - \eta(1-\lambda)\delta_i h_{i,1}^{-\lambda}[g^\lambda(y_i - \boldsymbol{\beta}'\mathbf{x}_i - \delta_i\eta) - C^{-1}\Delta] \\ 0 = \frac{(1-\delta_i)(1-\lambda)}{h_{i,0}} + \zeta(1-\delta_i) - \eta(1-\lambda)(1-\delta_i)h_{i,0}^{-\lambda}[g^\lambda(y_i - \boldsymbol{\beta}'\mathbf{x}_i) - C^{-1}\Delta] \end{cases},$$

or

$$\begin{cases} 0 = \delta_i(1-\lambda) + \zeta\delta_i h_{i,1} - \eta(1-\lambda)\delta_i h_{i,1}^{1-\lambda}[g^\lambda(y_i - \boldsymbol{\beta}'\mathbf{x}_i - \eta) - C^{-1}\Delta] \\ 0 = (1-\delta_i)(1-\lambda) + \zeta(1-\delta_i)h_{i,0} - \eta(1-\lambda)(1-\delta_i)h_{i,0}^{1-\lambda}[g^\lambda(y_i - \boldsymbol{\beta}'\mathbf{x}_i) - C^{-1}\Delta] \end{cases}. \tag{A.1}$$

Summing the above two equations together and over i, with the two constraints we get $\zeta = n(1-\lambda)$,

$$h_{i,1} = \frac{1}{n}\frac{1}{1+\eta h_i^{-\lambda}[g^\lambda(y_i - \boldsymbol{\beta}'\mathbf{x}_i - \eta) - C^{-1}\Delta]},$$

$$h_{i,0} = \frac{1}{n}\frac{1}{1+\eta h_i^{-\lambda}[g^\lambda(y_i - \boldsymbol{\beta}'\mathbf{x}_i) - C^{-1}\Delta]},$$

and η is determined by

$$\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{\delta_i}{1+\eta h^{-\lambda}g^\lambda(y_i - \boldsymbol{\beta}'\mathbf{x}_i - \eta)} + \frac{1-\delta_i}{1+\eta h^{-\lambda}g^\lambda(y_i - \boldsymbol{\beta}'\mathbf{x}_i)}\Big)^{1-\lambda}$$

$$\times[g^\lambda(y_i - \boldsymbol{\beta}'\mathbf{x}_i - \delta_i\eta) - C^{-1}\Delta] = 0.$$

Similarly as for empirical likelihood weights, assuming $g^\lambda(\epsilon_i)$ has second moment, then $\eta = O_p(n^{-1/2})$, and the equation for η is

$$0 = n^{-1}\sum_{i=1}^{n}\Big(\delta_i\big(1-\eta h^{-\lambda}g^\lambda(y_i - \boldsymbol{\beta}'\mathbf{x}_i - \eta)\big) + (1-\delta_i)\big(1-\eta h^{-\lambda}g^\lambda(y_i - \boldsymbol{\beta}'\mathbf{x}_i)\big)\Big)^{1-\lambda}$$

$$\times[g^\lambda(y_i - \boldsymbol{\beta}'\mathbf{x}_i - \delta_i\eta) - C^{-1}\Delta] + O_p(n^{-1})$$

$$= n^{-1}\sum_{i=1}^{n}\Big((1-\lambda)\delta_i\big(1-\eta h^{-\lambda}g^\lambda(y_i - \boldsymbol{\beta}'\mathbf{x}_i - \eta)\big) + (1-\lambda)(1-\delta_i)\big(1-\eta h^{-\lambda}g^\lambda(y_i - \boldsymbol{\beta}'\mathbf{x}_i)\big)\Big)$$

$$\times[g^\lambda(y_i - \boldsymbol{\beta}'\mathbf{x}_i - \delta_i\eta) - C^{-1}\Delta] + O_p(n^{-1}), \quad \text{or}$$

$$0 = n^{-1}\sum_{i=1}^{n}\Big(\delta_i\big(1-\eta h^{-\lambda}g^{\lambda}(y_i-\boldsymbol{\beta}'\mathbf{x}_i-\eta)\big)+(1-\delta_i)\big(1-\eta h^{-\lambda}g^{\lambda}(y_i-\boldsymbol{\beta}'\mathbf{x}_i)\big)\Big)$$
$$\times[g^{\lambda}(y_i-\boldsymbol{\beta}'\mathbf{x}_i-\delta_i\eta)-C^{-1}\Delta]+O_p(n^{-1})$$

and get, with $G(y_i-\boldsymbol{\beta}'\mathbf{x}_i-\delta_i\eta)=g^{\lambda}(y_i-\boldsymbol{\beta}'\mathbf{x}_i-\delta_i\eta)-C^{-1}\Delta$, let

$$\eta_0=\frac{\sum_{i=1}^{n}G(y_i-\boldsymbol{\beta}'\mathbf{x}_i-\delta_i\eta)}{\sum_{i=1}^{n}[\delta_i h^{-\lambda}g^{\lambda}(y_i-\boldsymbol{\beta}'\mathbf{x}_i-\eta)+(1-\delta_i)h^{-\lambda}g^{\lambda}(y_i-\boldsymbol{\beta}'\mathbf{x}_i)]G(y_i-\boldsymbol{\beta}'\mathbf{x}_i-\delta_i\eta)}.$$

We have

$$\eta=\frac{\sum_{i=1}^{n}G(y_i-\boldsymbol{\beta}'\mathbf{x}_i-\delta_i\eta)}{\sum_{i=1}^{n}[\delta_i h^{-\lambda}g^{\lambda}(y_i-\boldsymbol{\beta}'\mathbf{x}_i-\eta)+(1-\delta_i)h^{-\lambda}g^{\lambda}(y_i-\boldsymbol{\beta}'\mathbf{x}_i)]G(y_i-\boldsymbol{\beta}'\mathbf{x}_i-\delta_i\eta)}$$
$$+O_p(n^{-1}):=\eta_0+O_p(n^{-1}).$$

Plugging in the above expression for η into those for $h_{i,1}$ and $h_{i,0}$ we get their approximates.

Proof of Theorem 1. Recall the density $f(\cdot)=f(\cdot|\boldsymbol{\theta},h)$ given in (3). Let F be the distribution of f, $\hat{f}(\cdot)=f(\cdot|\hat{\boldsymbol{\theta}},\hat{\mathbf{h}})$, and $\hat{F}$ be the distribution function of $\hat{f}$. Denote $f_0(\cdot)=f(\cdot|\boldsymbol{\theta}_0,h_0)$. Let $\mathcal{B}$ be the Borel field on R^b, $H(\hat{f},f)$ be the Hellinger distance between $\hat{f}$ and f, and $||\hat{f}-f||$ be the variational distance, between $\hat{f}(\cdot)$ and $f(\cdot)$,

$$H(\hat{f},f)=2^{-1/2}\bigg[\int\Big(\hat{f}^{1/2}(\epsilon)-f^{1/2}(\epsilon)\Big)^2d\epsilon\bigg]^{1/2},$$
$$||\hat{f}-f||=2\sup\{|\hat{F}(B)-F(B)|:B\in\mathcal{B}\}=\int|\hat{f}(\epsilon)-f(\epsilon)|d\epsilon.$$

Recall the inequality $||\hat{f}-f||\leq 2H(\hat{f},f)$ (see Bickel et al.[25]). We will show that $H(\hat{f},f_0)\to 0$, a.s., so that $||\hat{f}-f_0||\to 0$, a.s., which implies $|\hat{h}_n(\cdot)-h_0(\cdot)|\to 0$ a.s. on $\mathcal{H}_n$ (defined below), and by condition (C4) $\sup_t|\hat{h}_n(t)-h_0(t)|\to 0$ a.s., and thus $\hat{f}_n(\cdot)\to f_0(\cdot)$, a.s., a.e. (L), with L being the Lebesque measure on R^d. Since the model is identifiable, we must have $\hat{\boldsymbol{\theta}}\to\boldsymbol{\theta}_0$ (a.s.), and get the desired result.

Below we show $H(\hat{f}, f) \to 0$ a.s.. For fixed $\boldsymbol{\theta}_0$ and $h_0 \in \mathcal{H}$, let $r_{\boldsymbol{\theta},h}(\cdot) = (\sqrt{f(\cdot|\boldsymbol{\theta},h)/f_0(\cdot)} - 1)1(f_0 > 0)$, $\mathcal{H}_n = \{h \in \mathcal{H} : h \text{ be of the form } h_{i,0}, h_{i,1}\}$ (as given in Section 2.2), $\mathcal{R}_n = \{r_{\boldsymbol{\theta},h} : \boldsymbol{\theta} \in \boldsymbol{\Theta}, h \in \mathcal{H}_n\}$, and $\mathcal{R} = \{r_{\boldsymbol{\theta},h} : \boldsymbol{\theta} \in \boldsymbol{\Theta}, h \in \mathcal{H}\}$. It is seen that $\mathcal{H}_n \subset \mathcal{H}$, $\mathcal{R}_n \subset \mathcal{R}$. Let P_n and P be empirical and the true distribution of the observed data. By Lemma 1.1 of van de Geer[26], since $(\hat{\boldsymbol{\theta}}, \hat{\mathbf{h}})$ is the semiparametric MLE of $(\boldsymbol{\theta}_0, h_0)$ on $\mathcal{R}_n$, based on model (2),

$$H^2(\hat{f}, f_0) \leq 2(P_n - P)\left(1(f_0 > 0)[\sqrt{\hat{f}/f_0} - 1]\right) = 2(P_n - P)r_{\hat{\boldsymbol{\theta}},\hat{h}}.$$

So to show $H(\hat{f}, f) \to 0$ a.s., it is suffice to show $\sup_{r\in\mathcal{R}_n} |(P_n - P)r| \to 0$ a.s., and since $\mathcal{R}_n \subset \mathcal{R}$, it suffice to show

$$\sup_{r\in\mathcal{R}} |(P_n - P)r| \to 0, \quad a.s.$$

i.e., $\mathcal{R}$ is a Glivenko-Cantelli class with respect to P.

For this, for a given probability measure P on $\mathcal{B}$, let $||g||_{L_1(P)} = \int |g(y)|P(dy)$, $N_{[\,]}(\epsilon, \mathcal{R}, L_1(P))$ be the minimum number of ϵ-brackets to cover $\mathcal{R}$ under norm $L_1(P)$, i.e. the minimum number k of pairs (l_j, u_j), $l_j, u_j \in \mathcal{R}$ such that $\forall r \in \mathcal{R}$, there is (l_j, u_j) $(1 \leq j \leq k)$ with $l_j \leq r \leq u_j$ and $||u_j - l_j||_{L_1(P)} \leq \epsilon$.

Below we need to evaluate $N_{[\,]}(\epsilon, \mathcal{R}, L_1(P))$. Let $\mathcal{F} = \{f_{\boldsymbol{\theta},h} : \boldsymbol{\theta} \in \boldsymbol{\Theta}, h \in \mathcal{H}\}$. Note that for all $f_1, f_2 \in \mathcal{F}$,

$$\left|\left|\left(\sqrt{\frac{f_1}{f_0}} - 1\right)1(f_0 > 0) - \left(\sqrt{\frac{f_2}{f_0}} - 1\right)1(f_0 > 0)\right|\right|_{L_1(P)}$$

$$= \int \frac{|f_1^{1/2}(\epsilon) - f_2^{1/2}(\epsilon)|}{f_0^{1/2}(\epsilon)} f_0(\epsilon)d\epsilon = \int |f_1^{1/2}(\epsilon) - f_2^{1/2}(\epsilon)| f_0^{1/2}(\epsilon)d\epsilon = C||\sqrt{f_1} - \sqrt{f_2}||_{L_1(Q)},$$

where C is some positive finite constant, Q is the probability measure corresponding to $\sqrt{f_0}$ (after normalization), and by condition (C2) this measure is well defined. Now, let $\mathcal{F}^{1/2} = \{\sqrt{f_{\boldsymbol{\theta},h}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}, h \in \mathcal{H}\}$. Since f_0 is fixed, the above equality gives $N_{[\,]}(\epsilon, \mathcal{R}, L_1(P)) \leq N_{[\,]}(\epsilon/C, \mathcal{F}^{1/2}, L_1(Q))$, for some $0 < C < \infty$.

Since by (C3), $\mathcal{F}^{1/2}$ is a collection of bounded continuous functions on $(R^+)^b$, so by Corollary 2.7.4 in van der Vaart and Wellner[27], with notations (V, d, α, r) there corresponds to $(b, b, 1, 1)$ here,

$$\log N_{[\,]}(\epsilon, \mathcal{F}^{1/2}, L_1(Q)) = O(\frac{1}{\epsilon^b}).$$

Thus, for some generic positive finite constant C,

$$N_{[\,]}(\epsilon, \mathcal{R}, L_1(P)) \leq N_{[\,]}(\frac{\epsilon}{C}, \mathcal{F}^{1/2}, L_1(Q)) \leq \exp\{C/\epsilon^b\} < \infty, \qquad \forall\ \epsilon > 0,$$

and so by Theorem 2.4.1 in van der Vaart and Wellner[27], $\mathcal{R}$ is a Glivenko-Cantelli class with respect to P, and complete the proof.

Proof of Theorem 2. For fixed $\boldsymbol{\theta}$, let $\hat{\mathbf{h}} = \hat{\mathbf{h}}(\boldsymbol{\theta})$ be the maximizer of likelihood (3) on $\mathcal{H}_n$, with $\langle_n$ given in the proof of Theorem 1. Note that (C5) implies $h_0(\cdot)$ is uniformly continuous, so as $n \to \infty$, $\hat{h}$ will be uniformly close to the global maximizer of h, and conditions (8)-(11) in Murphy and van der Vaart[22] can be satisfied, and by their Theorem 1, their expressions (4) and (5) hold, and their (5) gives the desired result, see also Proposition 2 in Severini and Wong[21], i.e.,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{D} N(\mathbf{0}, \Omega^{-1}(\boldsymbol{\theta}_0)), \qquad \Omega(\boldsymbol{\theta}_0) = E_{\boldsymbol{\theta}_0}[\mathbf{i}^* \mathbf{i}^{*'}],$$

where $\mathbf{i}^*$ is the efficient score for $\boldsymbol{\theta}_0$ under the original model (3). From this,

$$\sqrt{n}[(\hat{\boldsymbol{\beta}}', \hat{\eta})' - (\boldsymbol{\theta}_0', \eta_0)'] \xrightarrow{D} N(\mathbf{0}, \Omega_0^{-1}(\boldsymbol{\theta}_0)), \qquad \Omega_0(\boldsymbol{\theta}_0) = E_{\boldsymbol{\theta}_0, h_0}[\mathbf{i}_0^* \mathbf{i}_0^{*'}],$$

where $\mathbf{i}_0^*$ is the $(\boldsymbol{\beta}', \eta)'$ components of $\mathbf{i}^*$, $\mathbf{i}_0^*$ is the efficient score for $(\boldsymbol{\beta}_0', \eta_0)'$ under the original model (3). Theorem 1 in Murphy and van der Vaart[22] also requires some other conditions, such as the score function is P-Donsker over some neighborhood of the parameters, and the Hessian matrix is P-Glivenko-Cantelli over some neighborhood of the parameters. These conditions can be easily met under mild conditions, as

long as the class $\mathcal{H}$ of h's is regular. Checking conditions (8)-(11) in Murphy and van der Vaart is non-trival (checking conditions of Proposition 2 in Severini and Wong[21] may be no more easier, as they also require the consistency of derivative of $\hat{h}(\cdot)$ at some rate), but it can be followed by their lines for checking these conditions for the Cox model, in which the base line hazard function is maximized only at the observed data points, like our $\hat{h}$.

However, computation of $\mathbf{i}^*(\boldsymbol{\theta})$ is not easy, we only compute $I^*(\boldsymbol{\theta}_0)$. The log-likelihood for model (3) is

$$\ell(\boldsymbol{\theta},h)=\log\Big(\gamma g^{\lambda}(y-\boldsymbol{\beta}'\mathbf{x}-\eta)h^{1-\lambda}(y-\boldsymbol{\beta}'\mathbf{x}-\eta)+(1-\gamma)g^{\lambda}(y-\boldsymbol{\beta}'\mathbf{x})h^{1-\lambda}(y-\boldsymbol{\beta}'\mathbf{x})\Big)-\log\Delta.$$

Let $\dot{\ell}_0(\boldsymbol{\theta},h)=\partial\ell(\boldsymbol{\theta},h)/\partial(\boldsymbol{\beta}',\eta)'$, and $f(\epsilon|\boldsymbol{\theta},h)=\gamma g^{\lambda}(y-\boldsymbol{\beta}'\mathbf{x}-\eta)h^{1-\lambda}(y-\boldsymbol{\beta}'\mathbf{x}-\eta)+(1-\gamma)g^{\lambda}(y-\boldsymbol{\beta}'\mathbf{x})h^{1-\lambda}(y-\boldsymbol{\beta}'\mathbf{x})$, $\dot{g}(\epsilon)=\partial g(\epsilon)/\partial\epsilon$, and $\dot{h}(\epsilon)=\partial h(\epsilon)/\partial\epsilon$. Then

$$\begin{aligned}\dot{\ell}_0(\boldsymbol{\theta},h)=&-\Big(\gamma\lambda g^{\lambda-1}(y-\boldsymbol{\beta}'\mathbf{x}-\eta)h^{1-\lambda}(y-\boldsymbol{\beta}'\mathbf{x}-\eta)\dot{g}(y-\boldsymbol{\beta}'\mathbf{x}-\eta)\\&+\gamma(1-\lambda)g^{\lambda}(y-\boldsymbol{\beta}'\mathbf{x}-\eta)h^{-\lambda}(y-\boldsymbol{\beta}'\mathbf{x}-\eta)\dot{h}(y-\boldsymbol{\beta}'\mathbf{x}-\eta)\Big)/f(\epsilon|\boldsymbol{\theta},h)\mathbf{z}_1\\&-\Big((1-\gamma)\lambda g^{\lambda-1}(y-\boldsymbol{\beta}'\mathbf{x})h^{1-\lambda}(y-\boldsymbol{\beta}'\mathbf{x})\dot{g}(y-\boldsymbol{\beta}'\mathbf{x})\\&+(1-\gamma)(1-\lambda)g^{\lambda}(y-\boldsymbol{\beta}'\mathbf{x})h^{-\lambda}(y-\boldsymbol{\beta}'\mathbf{x})\dot{h}(y-\boldsymbol{\beta}'\mathbf{x})\Big)/f(\epsilon|\boldsymbol{\theta},h)\mathbf{z}_2\\:=&\ \mathbf{i}_{0,1}\mathbf{z}_1+\mathbf{i}_{0,2}\mathbf{z}_2\end{aligned}$$

where $\mathbf{z}_1=(\mathbf{x}',1)'$, $\mathbf{z}_2=(\mathbf{x}',0)'$.

Let ξ be the 0-1 valued random variable with $P(\xi=1)=\lambda$, η be the 0-1 valued random variable with $P(\eta=1)=\gamma$, and $q(\cdot)$ be the density-mass function (unknown) of $(\xi,\eta,\mathbf{z})$, $\dot{\mathbf{P}}_2$ and $\dot{\mathbf{P}}_3$ be the tangent space of $h(\cdot)$ and $q(\cdot)$, Λ be the nuisance space of (h,q), $\Lambda^{\perp}$ be its orthogonal complement, and $\Pi(\dot{\ell}|\Lambda)$ denote the projection of $\dot{\ell}$ onto Λ. Typically it is assumed that $h(\cdot)$ and $q(\cdot)$ are not functionally dependent each

other, then $\Lambda = \dot{\mathbf{P}}_2 \oplus \dot{\mathbf{P}}_3$, and so $\Pi(\dot{\ell}|\Lambda) = \Pi(\dot{\ell}|\dot{\mathbf{P}}_2) + \Pi(\dot{\ell}|\dot{\mathbf{P}}_3)$ here $\oplus$ means direct summation. The efficient score of $(\boldsymbol{\beta}', \eta')'$ in the presence of the nuisance parameters (h, q) is

$$\mathbf{i}_0^* = \Pi(\dot{\ell}_0|\Lambda^\perp) = \dot{\ell}_0 - \Pi(\dot{\ell}_0|\Lambda) = \dot{\ell}_0 - \Pi(\dot{\ell}_0|\dot{\mathbf{P}}_2) - \Pi(\dot{\ell}_0|\dot{\mathbf{P}}_3).$$

Let $\mathcal{B}_0$ be the σ field generated by ϵ, since $(\boldsymbol{\beta}', \eta')'$ is regression/location parameters, by (5.16) in Tsiatis[28] ,

$$\Pi(\dot{\ell}_0|\dot{\mathbf{P}}_2) = E\big(\dot{\ell}_0|\mathcal{B}_0\big) = E\big(\dot{\ell}_0|\epsilon\big) = \mathbf{i}_{0,1}(\epsilon)\boldsymbol{\mu}_1 + \mathbf{i}_{0,2}(\epsilon)\boldsymbol{\mu}_2,$$

where $\boldsymbol{\mu}_1 = E_{\boldsymbol{\theta}_0}[\mathbf{z}_1]$ and $\boldsymbol{\mu}_2 = E_{\boldsymbol{\theta}_0}[\mathbf{z}_2]$.

Also, $\dot{\ell}$ is orthogonal to $\dot{\mathbf{P}}_3$ (see, for example, Bickel et al.[25], Proposition 4.3.1 D and Example 4.3.1, p.104-105; or Tsiatis[28]), thus

$$\mathbf{i}_0^* = \dot{\ell}_0 - \Pi(\dot{\ell}_0|\dot{\mathbf{P}}_2) - \Pi(\dot{\ell}_0|\dot{\mathbf{P}}_3) = \dot{\ell}_0 - \Pi(\dot{\ell}_0|\dot{\mathbf{P}}_2) = \mathbf{i}_{0,1}(\epsilon)(\mathbf{z}_1 - \boldsymbol{\mu}_1) + \mathbf{i}_{0,2}(\epsilon)(\mathbf{z}_2 - \boldsymbol{\mu}_2)$$

and get $\Omega_0(\boldsymbol{\theta}_0) = E_{\boldsymbol{\theta}_0}[\mathbf{i}_0^*\mathbf{i}_0^{*'}]$.

Proof of the Proposition. Let $||f-g|| = \int |f(\epsilon)-g(\epsilon)|d\epsilon$, $C_0^{-1}(s) = \int g^s(\epsilon)h_0^{1-s}(\epsilon)d\epsilon$ be the normalizing constant for $g^s h_0^{1-s}$ and

$$\lambda = \arg\inf_{s\in[0,1]} ||C_0(s)g^s h_0^{1-s} - [\rho g + (1-\rho)h_0]||, \quad \delta = ||C_0(\lambda)g^\lambda h_0^{1-\lambda} - [\rho g + (1-\rho)h_0]||.$$

If $\delta = 0$, there is nothing to prove. If $\delta > 0$, we divide the integration region S as $S = S_+ \cup S_-$, with $S_+ = \{\epsilon : C_0(\lambda)g^\lambda(\epsilon)h_0^{1-\lambda}(\epsilon) \geq \rho g(\epsilon) + (1-\rho)h_0(\epsilon)\}$ and $S_- = S \setminus S_+$, then both S_+ and S_- are non-empty. It is clear that there is a function $r(\cdot)$, such that $0 \leq r^{1-\lambda}(\cdot) \leq 1$ on S_+ and $r^{1-\lambda}(\cdot) > 1$ on S_-, and

$$C_0(\lambda)g^\lambda(\cdot)h_0^{1-\lambda}(\cdot)r^{1-\lambda}(\cdot) \equiv \rho g(\cdot) + (1-\rho)h_0(\cdot).$$

i.e. $r(\cdot)$ downward adjusts $C_0(\lambda)g^\lambda(\cdot)h_0^{1-\lambda}(\cdot)$ on S_+ and upward adjusts it on S_-, to achieve the equality with $\rho g(\cdot) + (1-\rho)h_0(\cdot)$. Specifically, $r(\cdot)$ is given by

$$r^{1-\lambda}(\epsilon) = \frac{\rho g(\epsilon) + (1-\rho)h_0(\epsilon)}{C_0(\lambda)g^\lambda(\epsilon)h_0^{1-\lambda}(\epsilon)} I\big\{C_0(\lambda)g^\lambda(\epsilon)h_0^{1-\lambda}(\epsilon) > 0\big\}.$$

Now let $h(\cdot) = ch_0(\cdot)r(\cdot)$, $c^{-1} = \int h_0(\epsilon)r(\epsilon)d\epsilon$, then h is a density, and replace $C_0(\lambda)$ by $C(\lambda) = C_0(\lambda)c^{1-\lambda}$, and $C_0(\lambda)g^\lambda(\cdot)[h_0(\cdot)r(\cdot)]^{1-\lambda} = C(\lambda)g^\lambda(\cdot)h^{1-\lambda}(\cdot)$, and we have

$$C(\lambda)g^\lambda(\cdot)h^{1-\lambda}(\cdot) \equiv \rho g(\cdot) + (1-\rho)h_0(\cdot).$$

The statement about the correspondence between ρ and λ is obvious.

REFERENCES

1. Goldhrisch ASC, Collieoni M, Nasi ML, Bernhard J, Zahrieh D, Bonetti M, Gelber RD, Italiana S, Bellinzona S, Coates AS, et al. Endocrine responsiveness and tailoring adjuvant therapy for postmenopausal lymph node-negative breast cancer: a randomized trial. *Journal of the National Cancer Institute*. 2002; **94**:10541065.

2. Sabine C. AIDS events among individuals initiating HAART: do some patients experience a greater benefit from HAART than others? *AIDS*. 2005;**19**:19952000.

3. Bonetti M, Gelber RD. Patterns of treatment effects in subsets of patients in clinical trials. *Biostatistics*. 2005;**5**:465481.

4. Ruberg SJ, Chen L, Wang Y. The mean doesn't mean as much anymore: finding sub-groups for tailored therapeutics, *Clinical Trials*. 2010;**7**:574-583.

5. Cai T, Tian L, Wong P, Wei LJ. Analysis of randomized comparative clinical trial data for personalized treatment selections.*Biostatistics*. 2011; **12**:270-282.

6. Shen J, He X. Inference for subgroup analysis with a structured logistic-normal mixture model, *Journal of the American Statistical Association*. 2015;**110**:303-312.

7. Rothmann MD, Zhang J, Lu L, Fleming TR. Testing in a pre-specified subgroup and the intent-to-treat population. *Drug Information Journal*. 2012;**46**(2):175-179.

8. Friede T, Parsons N, Stallard N. A conditional error function approach for subgroup selection in adaptive clinical trials. *Statistics in Medicine*. 2012;**31**:4309-4320.

9. Song Y, Chi GY. A method for testing a pre-specified subgroup in clinical trials. *Statistics in Medicine*. 2007;**26**:3535-3549.

10. Alosh M, Huque MF. A flexible strategy for testing subgroups and overall population. *Statistics in Medicine*. 2009;**28**:3-23.

11. Sivaganesan S, Laudm PW, Muller P. A Bayesian subgroup analysis with a zero-enriched Polya urn scheme. *Statistics in Medicine*. 2011;**30**:312-323.

12. Jones HE, Ohlssen DI, Neuenschwander B, Racine A, Branson M. Bayesian models for subgroup analysis in clinical trials. *Clinical Trials*. 2011;**8**:129-143.

13. Su X, Tsai CL, Wang H, Nickerson DM, Li B. Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*. 2009;**10**:141-158.

14. Lipkovich I, Dmitrienko A, Denne J, Enas G. Subgroup identification based on differential effect search (SIDES) A recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in Medicine*. 2011;**30**:2601-2621.

15. Foster JC, Taylor JMC, Ruberg SJ. Subgroup identification from randomized clinical trial data, *Statistics in Medicine*. 2011;**30**:2867-2880.

16. Olkin I. A semiparametric approach to density estimation. *Journal of the American Statistical Association*. 1987; **82**: 858-865.

17. Vardi Y. Empirical distributions in selection bias models, *Annals of Statistics*. 1985;**13**:178-203.

18. Qin J. Empirical likelihood in biased sample problems. *Annals of Statistics*. 1993; **21**:1182-1196.

19. Gilbert P. Large sample theory of maximum likelihood estimates in semiparametric biased sampling models, *Annals of Statistics*. 2000;**28**:151-194.

20. Altstein, Li. Latent subgroup analysis of a randomized clinical trial through a semiparametric accelerated failure time mixture model.*Biometrics*. 2013;**69**(1):52-61.

21. Severini TA,Wong WH. Profile likelihood and conditionally parametric models. *Annals of Statistics*. 1992; **20**:1768–1802.

22. Murphy SA, Van der Vaart AW. On profile likelihood. *Journal of the American Statistical Association*. 2000; **93**:1461–1474.

23. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society*. Ser. B, 1977;**39**:1-38.

24. Tan M, Tian GL, Ng KW. *Bayesian Missing Data Problems: EM, Data Augmentation and Non-iterative Computation*, London and Boca Raton, Florida: Chapman and Hall/CRC,2009.

25. Bickel PJ, Klaassen CA, Ritov Y, Wellner JA. *Efficient and Adaptive Estimation for Semiparametric Models*, Johns Hopkins University Press, Baltimore, Maryland,1993.

26. van de Geer S. Hellinger-consistency of certain nonparametric maximum likelihood estimates, *Annals of Statistics*. 1993;**21**:14-44.

27. van der Vaart A, Wellner J. *Weak Convergence and Empirical Processes*, Springer;1996 .

28. Tsaitis AA. *Semiparametric Theory and Missing Data*, Springer, New York, 2006.

29. van der Vaart. *Semiparametric Statistics*, in Part III, *Lectures on Probability Theory and Statistics*, Springer; 1999.

30. Gao, Z., Roy, A., Tan, M. A two-stage adaptive targeted clinical trial design for biomarker performance- based sample size re-estimation. *Statistics in Biosciences, Journal of the International Chinese Statistical Association*. Springer, 2016. DOI 10.1007/s12561-015-9139-3.

Email: Xiaofei Chen, xc81@georgetown.edu; Ao Yuan, ay312@georgetown.edu; Yizhao Zhou, yz459@georgetown.edu; Ming T. Tan, mtt34@georgetown.edu.