UW Biostatistics Working Paper Series

2-18-2009

Semiparametric Two-Part Models with Proportionality Constraints: Analysis of the Multi-Ethnic Study of Atherosclerosis (MESA)

Anna Liu
University of Massachusetts

Richard Kronmal
University of Washington, kronmal@u.washington.edu

Xiao-Hua Zhou
University of Washington, azhou@u.washington.edu

Shuangge Ma
Yale University, shuangge.ma@yale.edu

Suggested Citation

Liu, Anna; Kronmal, Richard; Zhou, Xiao-Hua ; and Ma, Shuangge, "Semiparametric Two-Part Models with Proportionality Constraints: Analysis of the Multi-Ethnic Study of Atherosclerosis (MESA)" (February 2009). *UW Biostatistics Working Paper Series.* Working Paper 341.
http://biostats.bepress.com/uwbiostat/paper341

Semiparametric Two-Part Models with Proportionality Constraints: Analysis of The Multi-Ethnic Study of Atherosclerosis (MESA)

Anna Liu[1], Richard Kronmal[2], Xiaohua Zhou[2,3] and Shuangge Ma[4*]

[1]Department of Mathematics and Statistics, University of Massachusetts

[2]Department of Biostatistics, University of Washington

[3]Biostatistics Unit, HSR&D Center of Excellence Veterans Affairs Puget Sound Health Care System

[4]Department of Epidemiology and Public Health, Yale University

[*]*email:* shuangge.ma@yale.edu

Summary. In this article, we analyze the coronary artery calcium (CAC) score in the Multi-Ethnic Study of Atherosclerosis (MESA), where about half of the CAC scores are zero and the rest are continuously distributed. When the observed data has a mixture distribution, two-part models can be the natural choice. With a two-part model, there are two covariate effects, with one in each part of the model. Determination of whether the two covariate effects are proportional can provide more insights into the process underlying development and progression of CAC. In this study, we model the CAC score using a semiparametric two-part model, and investigate the determination of proportionality of the covariate effects. We propose penalized maximum likelihood estimation and using thin plate splines in practical data analysis, and establish asymptotic estimation properties. We propose a step-wise hypothesis testing based approach to determine proportionality. Simulation studies suggest satisfactory finite-sample performance of the proposed approach. Analysis of the MESA data suggests that proportionality holds for all covariates except the LDL and HDL.

Key words: Two-part models; Proportionality; Semiparametric estimation; Splines.

1. Introduction

Statistical development in this article has been motivated by analysis of the coronary artery calcium (CAC) in the MESA (Multi-Ethnic Study of Atherosclerosis). The MESA is an ongoing study of the prevalence, risk factors, and progression of subclinical cardiovascular disease in a multi-ethnic cohort (Bild et al. 2002). In previous studies, the CAC has been established as an important risk factor for the development of various coronary heart diseases. Understanding the development of CAC can be valuable for clinical diagnosis and treatment of multiple cardiovascular diseases. In the MESA, the CAC is measured with the Agatston score, which is the amount of calcium at each lesion scaled by an attenuation factor and summed over all lesions. We show the histogram of $\log(1 + CAC)$ in Figure 1. It is clear that, the CAC has a mixture distribution: about half of the CAC scores are zero, and the rest are continuously distributed.

In biomedical studies, data with mixture distributions are commonly encountered. Denote Y as the response variable of interest. In this article, we consider a special form of mixture distributions: for a subset of subjects, $Y = c$ with a fixed c; and for the rest of the subjects, $Y \sim \xi(Y)$ where ξ is a continuous density function. Methodologies developed in this article are applicable to other mixture distributions with minor modifications. When responses with mixture distributions are observed, two-part models can be the natural choice. Two-part models have a long history in economic, statistical, and biomedical literature. On a special note, two-part models have been suggested as the default models for describing the CAC in MESA (http://mesa-nhlbi.org/).

For the type of data described above, we consider the following two-part models. Denote $X = (X_1, X_2, X_3)$ as the covariate. In the first part of the model, we assume

$$\phi^{-1}(Pr(Y = c|X)) = h(X), \tag{1}$$

where ϕ is a known monotone (increasing) transformation function, ϕ^{-1} is the inverse of ϕ,

and $h(X)$ is the unknown covariate effect. In the second part of the model, we assume

$$\text{for } Y \neq c: \ Y|X = h^*(X) + \epsilon, \tag{2}$$

where $h^*(X)$ is the unknown covariate effect and ϵ is the random error with a known distribution. If $h^*(X) = \tau h(X)$ with $\tau \neq 0$, we conclude that *the two covariate effects are proportional*. When the proportionality does not hold, there can be multiple scenarios. Consider for example the additive covariate effects, where $h(X) = h_1(X_1) + h_2(X_2) + h_3(X_3)$. If $h^*(X) = \tau(h_1(X_1) + h_2(X_2) + h_3(X_3)) + \tilde{h}(X_1)$ with $\tilde{h}(X_1) \neq 0$ and $\tau \neq 0$, we conclude *partial proportionality*. That is, proportionality (of covariate effects) holds for X_2 and X_3, but not for X_1. Other partial proportionality scenarios can be defined in a similar manner. For simplicity of notations, we assume three covariates. Proportionality can be defined accordingly when there are more or fewer covariates.

Determination of proportionality with two-part models can be of critical interest. For the CAC, if the covariate effects are proportional, then the same function of the predictors determines if the CAC is zero as well as its actual level if nonzero. Such a result, if obtained, can confirm the hypothesis that the change from a zero to a positive Agatston score and the change from a lower to a higher Agatston score share the same underlying biological process. If partial proportionality can be obtained, then covariates can be naturally separated into two groups: proportional and non-proportional ones. Most likely, the two groups of covariates determine the CAC levels via two separate processes.

Determination of proportionality with two-part models can be traced back to Cragg (1971). Other examples include the zero-inflated Poisson regression model in Lambert (1992) and Albert et al. (1997), and the logit-(log) gamma two part model in Moulton et al. (2002). The most closely related study is Han and Kronmal (2006), where determination of proportionality with parametric two-part models is investigated. Published studies suggest that determination of proportionality can provide more insights into the biological mechanisms

underlying (for example) disease developments. In addition, compared with models without proportionality constraints, models with proportional (or partially proportional) covariate effects have fewer unknown parameters and thus can be more accurately estimated. A common drawback of the aforementioned studies is that, parametric models with strong assumptions have been used.

Semiparametric two-part models may be needed beyond parametric models. McClelland et al. (2006) studied the CAC in MESA and showed that certain covariate effects are nonlinear. Semiparametric two-part models have been investigated in recent years. Examples include Lam and Xue (2005), Ma (2009), and references therein, where semiparametric models for the density function ξ and covariate effect h^* are considered. In those studies, the focus has been semiparametric estimation and the forms of covariate effects have been assumed to be known. With semiparametric two-part models, we expect that determination of proportionality with respect to parametric covariate effects can be achieved using likelihood-based hypothesis testing approaches, although such an aspect has not been investigated. On the other hand, it is not clear how to determine proportionality with respect to nonparametric covariate effects.

In this article, for semiparametric two-part models, we investigate determination of proportionality of covariate effects. Our study has been motivated by analysis of the CAC in MESA, although the proposed methodology is applicable to many other mixture distributions and other two-part models. Methodological development in this article contains two major components: development of a hypothesis testing based approach for determining proportionality and establishment of asymptotic estimation properties.

This study advances from published literature along the following aspects. First, compared with existing proportionality studies of parametric two-part models, more flexible semiparametric models are adopted, which can provide better descriptions of data. Second,

the proposed hypothesis testing approach for determining proportionality advances from published studies by studying more complicated semiparametric models, and adopting a step-wise method that can accommodate multiple nonparametric and parametric covariate effects. Third, this study advances from published analysis of semiparametric two-part models by investigating different models, rigorously establishing asymptotic properties, and more importantly proposing an effective approach for determining proportionality. Last, a more comprehensive analysis of the CAC is conducted, which can provide a deeper understanding of the development of CAC and coronary heart diseases.

The rest of the article is organized as follows. The data and model setting is introduced in Section 2. The proposed methodology is described in Section 3. We consider penalized maximum likelihood estimation, and use thin plate splines with finite-sample data. We propose a hypothesis testing approach for determination of proportionality, and establish asymptotic estimation properties. Simulation studies are presented in Section 4. We analyze the MESA data in Section 5. The article concludes with discussions in Section 6. Proofs are provided in the Appendix.

2. Data and Model

Denote Y as the response of interest, where with a nonzero probability $Y = c$. In the analysis of CAC, $Y = \log(1 + CAC)$ and $c = 0$. Denote $X = (X_1, X_2, X_3)'$ and $Z = (Z_1, Z_2, Z_3)'$ as covariates. The proposed methodology is straightforwardly applicable when there are a different number of covariates.

In the first part of the two-part model, we assume that

$$\begin{aligned} \phi^{-1}(Pr(Y = 0|X, Z)) &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + f_1(Z_1) + f_2(Z_2) + f_3(Z_3) \\ &= \beta' \tilde{X} + f(Z), \end{aligned} \tag{3}$$

where ϕ is the known link function and ϕ^{-1} is the inverse of ϕ. Multiple link functions are

available, with the logit link most extensively used. $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)'$, $\tilde{X} = (1, X')'$, and $f(Z) = f_1(Z_1) + f_2(Z_2) + f_3(Z_3)$. In the second part of the model, we assume that for $Y \neq 0$

$$\begin{aligned} Y|X, Z &= \tau(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + f_1(Z_1) + f_2(Z_2) + f_3(Z_3)) \\ &\quad + \alpha_0 + \alpha_2 X_2 + \alpha_3 X_3 + g_1(Z_1) + g_2(Z_2) + g_3(Z_3) + \boldsymbol{\epsilon} \\ &= \tau(\beta' \tilde{X} + f(Z)) + \alpha' \tilde{\tilde{X}} + g(Z) + \boldsymbol{\epsilon} \end{aligned} \tag{4}$$

where $\alpha = (\alpha_0, \alpha_2, \alpha_3)'$, $\tilde{\tilde{X}} = (1, X_2, X_3)'$, $g(Z) = g_1(Z_1) + g_2(Z_2) + g_3(Z_3)$, and $\boldsymbol{\epsilon}$ has a known distribution. Motivated by Figure 1, we assume $N(0, \sigma^2)$ distributed error with unknown σ.

In (3) and (4), α, β, τ and σ are the unknown parametric regression parameters. f and g are the unknown nonparametric covariate effects. For simplicity of notations, we assume additive covariate effects, which can be easily extended to more general cases. Motivated by the findings in McClelland et al. (2006), we assume f and g are smooth functions.

In (4), proportionality holds if $\alpha = 0$ and $g = 0$. Thus, determination of proportionality (or partial proportionality) amounts to testing whether α and g (or their components) are equal to zero. For identifiability, X_1 is not included in $\tilde{\tilde{X}}$, and will be referred to as the "anchor" covariate. We also assume that $\tau\beta_1 \neq 0$.

3. Penalized Estimation and Determination of Proportionality

3.1 *Penalized estimation*

For an observation with covariate (X, Z) and response Y, the log-likelihood function is

$$\begin{aligned} & l(\alpha, \beta, \tau, \sigma, f, g|X, Z) \\ = \; & I(Y \neq 0) \left\{ -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{(Y - \tau(\beta' \tilde{X} + f(Z)) - \alpha' \tilde{\tilde{X}} - g(Z))^2}{2\sigma^2} \right\} \\ & + I(Y \neq 0)\log(1 - \phi(\beta' \tilde{X} + f(Z))) + I(Y = 0)\log(\phi(\beta' \tilde{X} + f(Z))). \end{aligned} \tag{5}$$

In what follows, we set ϕ as the logit link function. Assume there are n iid observations. Under the assumption of smooth f and g, we consider the penalized maximum likelihood

estimate (PMLE)

$$(\hat{\alpha}, \hat{\beta}, \hat{\tau}, \hat{\sigma}, \hat{f}, \hat{g}) = argmax \left\{ P_n l - \lambda_f^2 J^2(f) - \lambda_g^2 J^2(g) \right\}, \tag{6}$$

where P_n is the empirical measure, λ_f and λ_g are the data-dependent tuning parameters, J is the penalty on smoothness defined as $J^2(f) = \sum_{i=1}^{3} J^2(f_i) = \sum_{i=1}^{3} \int (f_i^{(s)})^2 dZ_i$, and $f_i^{(s)}$ is the s^{th} derivative of f_i.

Penalized estimation has been extensively used with semiparametric models when unknown smooth functions are present. An advantage of penalization estimation is that the smoothness of estimates is directly controlled by the data-dependent tuning parameters. We note that, other smoothing techniques, such as the local polynomials, can also be used.

3.2 *Finite-sample estimation with thin plate splines*

As shown in the Appendix, under assumptions described in Section 3.4, $\hat{f}$ and $\hat{g}$ are splines. In practice, with finite-sample data, we estimate f and g with thin plate splines. For a generic function $m(x)$, its thin plate spline representation is

$$m(x) = d_0 + d_1 x + \sum_{k=1}^{K} c_k |x - p_k|^3, \tag{7}$$

where d_0, d_1 and c_ks are the unknown regression coefficients and p_ks are the fixed knots.

For $i = 1, 2, 3$, at the design points, we have

$$f_i(Z_i) = T_i \mathbf{d}_{fi} + \Sigma_i \mathbf{c}_{fi}, \qquad g_i(Z_i) = T_i \mathbf{d}_{gi} + \Sigma_i \mathbf{c}_{gi},$$

where $T_i = (1, Z_i)$, $\Sigma_i = (|Z_i - p_{i1}|^3, \cdots, |Z_i - p_{iK}|^3)$, p_{ik}s are the knots, and $\mathbf{d}_{fi} = (d_{0fi}, d_{1fi})'$, $\mathbf{d}_{gi} = (d_{0gi}, d_{1gi})'$, $\mathbf{c}_{fi} = (c_{1fi}, \cdots, c_{Kfi})'$, $\mathbf{c}_{gi} = (c_{1gi}, \cdots, c_{Kgi})'$ are the regression coefficients. In our study, selection of knots follows Wahba (1990). Denote

$$\boldsymbol{\theta} = (\alpha_0, \alpha_2, \alpha_3, \beta_0, \beta_1, \beta_2, \beta_3, \tau, \sigma, \mathbf{d}_{f1}', \mathbf{d}_{f2}', \mathbf{d}_{f3}', \mathbf{d}_{g1}', \mathbf{d}_{g2}', \mathbf{d}_{g3}')'$$

and $\mathbf{b} = (\mathbf{c}_{f1}', \mathbf{c}_{f2}', \mathbf{c}_{f3}', \mathbf{c}_{g1}', \mathbf{c}_{g2}', \mathbf{c}_{g3}')'$. With the proposed penalized estimation, once the knots are chosen, penalization on the smoothness (i.e., $J(f)$ and $J(g)$) is equivalent to penalization

on the coefficients $\mathbf{b}$. In practical data analysis, instead of using unified λ_f and λ_g for all components of f and g, we can use different λ_{fi} and λ_{gi} for $i = 1, 2, 3$. With these notations, the penalized log-likelihood function defined in (6) can be rewritten as

$$P_n l(Y|\boldsymbol{\theta}, \mathbf{b}, \sigma^2) - \sum_{i=1}^{3} \lambda_{fi}^2 \mathbf{c}_{fi}' D_i \mathbf{c}_{fi} - \sum_{i=1}^{3} \lambda_{gi}^2 \mathbf{c}_{gi}' D_i \mathbf{c}_{gi} \tag{8}$$

with $l(Y|\boldsymbol{\theta}, \mathbf{b}, \sigma^2) = I(Y = 0)\boldsymbol{\eta}_1 - \log(1+\exp(\boldsymbol{\eta}_1)) - I(Y \neq 0)\left(\frac{1}{2}\log(2\pi) + \frac{1}{2}\log\sigma^2 + \frac{(Y-\boldsymbol{\eta}_2)^2}{2\sigma^2}\right)$, $\boldsymbol{\eta}_1 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \sum_{i=1}^{3}(T_i \mathbf{d}_{fi} + \Sigma_i \mathbf{c}_{fi})$, $\boldsymbol{\eta}_2 = \tau\boldsymbol{\eta}_1 + \alpha_0 + \alpha_2 X_2 + \alpha_3 X_3 + \sum_{i=1}^{3}(T_i \mathbf{d}_{gi} + \Sigma_i \mathbf{c}_{gi})$, and $D_i = (|p_{ik} - p_{i1}|^3, \cdots, |p_{ik} - p_{iK}|^3)_{k=1}^{K}$.

Since the objective function defined in (8) is concave in both $\boldsymbol{\theta}$ and $\mathbf{b}$, maximization can be achieved simply using the Newton-Raphson algorithm.

3.2.1 Tuning parameter selection For selection of the optimal tuning parameters, we propose a Generalized Maximum Likelihood (GML) smoothing parameter selection approach, which has been motivated by the approach developed in Wahba (1990) for Gaussian data. The GML criterion considers (8) as the joint likelihood of the response Y and the following random effects:

$$\mathbf{c}_{fi} \sim N(\mathbf{0}, D_i^+/\lambda_{fi}^2), \mathbf{c}_{gi} \sim N(\mathbf{0}, D_i^+/\lambda_{gi}^2), \quad i = 1, 2, 3, \tag{9}$$

where D_i^+ is the Moore-Penrose inverse of D_i.

If we assume a flat prior on $\boldsymbol{\theta}$, the GML criterion then estimates the smoothing parameters and σ^2 from the marginal density of Y, which is

$$\begin{aligned} & L(Y|\lambda_{f1}, \lambda_{f2}, \lambda_{f3}, \lambda_{g1}, \lambda_{g2}, \lambda_{g3}, \sigma^2) \\ = & \int \exp\left(P_n l(Y|\boldsymbol{\theta}, \mathbf{b}, \sigma^2) - \sum_{i=1}^{3} l(\mathbf{c}_{fi}) - \sum_{i=1}^{3} l(\mathbf{c}_{gi})\right) d\boldsymbol{\theta} d\mathbf{c}_{f1} \cdots d\mathbf{c}_{g3}, \end{aligned} \tag{10}$$

where $l(\mathbf{c}_{fi})$ and $l(\mathbf{c}_{fi})$ are the log-likelihood functions of the normal distributions in (9).

If $l(Y|\boldsymbol{\theta}, \mathbf{b}, \sigma^2)$ were a normal likelihood, the GML criterion gives the REML estimates of the tuning parameters, which are the inverse of the variance components in a mixed effects model with $\mathbf{c}_{fi}$s and $\mathbf{c}_{gi}$s as the random effects. Under this mixed effects model framework, alternatively, we can use a full marginal likelihood (ML) approach, which allows us to estimate the fixed effect $\boldsymbol{\theta}$ together with the variance components. Here, the full marginal likelihood of Y is

$$\begin{aligned} & L(Y|\boldsymbol{\theta}, \lambda_{f1}, \lambda_{f2}, \lambda_{f3}, \lambda_{g1}, \lambda_{g2}, \lambda_{g3}, \sigma^2) \\ = & \int \exp\left(P_n l(Y|\boldsymbol{\theta}, \mathbf{b}, \sigma^2) - \sum_{i=1}^{3} l(\mathbf{c}_{fi}) - \sum_{i=1}^{3} l(\mathbf{c}_{gi}) \right) d\mathbf{c}_{f1} \cdots d\mathbf{c}_{g3}. \end{aligned} \tag{11}$$

Of note, the REML and ML approaches are asymptotically equivalent, with the former more efficient for estimating the variance components, and the latter more convenient for inferences involving fixed effects. In this study, since estimation and testing of both fixed effects and tuning parameters are of interest, the ML approach is adopted.

Numerically, we carry out the multivariate integration in (11) using the spherical-radial quadrature algorithm, which is proposed by Monohan and Genz (1997) in the context of Bayesian computation. For generalized linear mixed effects models, Clarkson and Zhan (2002) showed that the spherical-radial multiple integration algorithm performs better than the second-order Laplace approximations. In addition, it is computationally more affordable than the Bayes sampling, while having comparable performances.

3.3 *Determination of proportionality*

Determination of proportionality with respect to X_i is equivalent to testing

$$H_0 : \alpha_i = 0 \quad vs \quad H_1 : \alpha_i \neq 0, \qquad i = 2, 3.$$

With Z_i, determination of proportionality amounts to testing

$$H_0 : \mathbf{d}_{gi} = \mathbf{0}, \lambda_{gi} = \infty \quad vs \quad H_1 : \mathbf{d}_{gi} \neq \mathbf{0} \text{ or } \lambda_{gi} \neq \infty, \qquad i = 1, 2, 3.$$

When there is a single nonparametric covariate effect, similar hypothesis testing problems have been considered for linear models (Wahba 1990) and generalized linear models (Liu et al. 2005). The aforementioned studies demonstrate satisfactory performance of likelihood ratio based approaches. Motivated by those studies as well as Guo (2002) and Crainiceanu et al. (2005), for both parametric and nonparametric covariate effects, we propose using the following likelihood ratio test statistic based on the ML defined in (11):

$$T_{ML} = \frac{\sup_{H_0} L(Y|\boldsymbol{\theta}, \lambda_{f1}, \lambda_{f2}, \lambda_{f3}, \lambda_{g1}, \lambda_{g2}, \lambda_{g3}, \sigma^2)}{\sup_{H_0 \cup H_1} L(Y|\boldsymbol{\theta}, \lambda_{f1}, \lambda_{f2}, \lambda_{f3}, \lambda_{g1}, \lambda_{g2}, \lambda_{g3}, \sigma^2)}. \tag{12}$$

In our study, there are are multiple covariates, and multiple different scenarios of partial proportionality. To fully determine the proportionality property, we consider the following forward step-wise approach. Denote A, A_P and A_N as the index sets of all covariates, covariates with proportional effects, and covariates with non-proportional effects, respectively. Denote C_{AP} as the cardinality of A_P.

1. Initialize $A_P = A$;

2. For $\forall a \in A_P$, fit an intermediate model, where covariates with index in $A_P - \{a\}$ have proportional effects, and covariates with index in $A_N \cup \{a\}$ have non-proportional effects. Compute the p-value for proportionality using the bootstrap approach described below.

3. Repeat Step 2 over all $a \in A_P$, and compare the C_{AP} p-values so obtained. Denote a^* as index of the covariate with the smallest p-value. If the smallest p-value is not significant, abort loop. Otherwise, update A_P with $A_P - \{a^*\}$ and A_N with $A_N \cup \{a^*\}$.

4. If $C_{AP} = 0$, abort loop. Otherwise, iterate Steps 2 and 3.

This step-wise approach starts with all covariate effects being proportional. In Step 2, we only need to determine significance of proportionality with respect to one covariate effect

at a time. With Step 3, at each iteration, the proportionality constraint on one covariate effect is released. Iteration is terminated once A_P cannot be further reduced.

For the parametric regression parameters, in theory, hypothesis testing can be based on the asymptotic normality result established in Section 3.4 and a variance estimate. However, our investigation shows that the asymptotic variance does not have a simple analytic form. For the nonparametric covariate effects, Liu et al. (2005) showed that for models much simpler than the proposed ones, bootstrap is needed for hypothesis testing. To compute the p-values of proportionality, we propose the following bootstrap approach.

1. Fit the Null model;
2. With observed covariate values, generate random errors from the normal distribution with mean zero and variance $\hat{\sigma}^2$;
3. Generate the binary $I(Y \neq 0)$ using model (3) and dichotomizing the probabilities at 0.5; For those with $Y \neq 0$, generate the continuous Y values under the null model;
4. With the generated responses, estimate the model again; Compute the statistic T_{ML};
5. Repeat Steps 2 to 5 B (e.g. 500) times. An empirical p-value can then be computed.

The proposed bootstrap approach shares similar spirits with Liu et al. (2005). We investigate its empirical performance in Section 4. We note that, a byproduct of the above procedure is the bootstrap confidence intervals for both the parametric and nonparametric parameters, which can serve as basis for inference.

3.4 *Asymptotic estimation properties*

Although many intermediate models need to be fit in order to determine the proportionality property, we are most interested in the "final models", i.e., models with proportionality

properly determined. In this section, for the final models, we establish asymptotic properties of the PMLE defined in (6). First, we make the following assumptions.

(A1) The covariates X and Z are component-wise bounded. The true value of $(\alpha, \beta, \tau, \sigma)$, denoted as $(\alpha_T, \beta_T, \tau_T, \sigma_T)$, is an interior point of a compact set.

(A2) Denote f_T and g_T as the unknown true values of f and g, respectively. Component-wise, f_T and g_T belong to the Sobolev space indexed by the order of derivative s. In this study, we adopt the commonly assumed $s = 2$. For identifiability, we also assume $Pf_i = Pg_i = 0$, where P is the expectation.

(A3) Define $d^2((\alpha, \beta, \tau, \sigma, f, g), (\alpha_T, \beta_T, \tau_T, \sigma_T, f_T, g_T)) = |\alpha - \alpha_T|^2 + |\beta - \beta_T|^2 + |\tau - \tau_T|^2 + |\sigma - \sigma_T|^2 + \int (f - f_T)^2 dX + \int (g - g_T)^2 dZ$. Assume $P(l(\alpha, \beta, \tau, \sigma, f, g) - l(\alpha_T, \beta_T, \tau_T, \sigma_T, f_T, g_T)) \leq -K_1 d^2((\alpha, \beta, \tau, \sigma, f, g), (\alpha_T, \beta_T, \tau_T, \sigma_T, f_T, g_T))$ with a fixed constant $K_1 > 0$.

(A4) $\lambda_f, \lambda_g = O_p(n^{-s/(2s+1)})$.

For most practical data, the boundedness assumption A1 is satisfied. We make this assumption for theoretical convenience only, and allow the actual bounds to remain unknown. We assume the nonparametric covariate effects are spline functions in A2. We assume the maximizer of the likelihood function is "well-separated" in A3. This assumption can be satisfied under the boundedness assumptions A1 and A2 and the differentiability of the likelihood function. If the tuning parameters λ_f, λ_g have the order as assumed in A4, the optimal convergence rate can be obtained, as shown below. In practice, they will be chosen using the approach described in Section 3.2.1.

For the final models, asymptotic properties of the PMLE can be summarized in the following two lemmas.

Lemma 1. *Under assumptions A1-A4,*

$$d((\hat{\alpha}, \hat{\beta}, \hat{\tau}, \hat{\sigma}, \hat{f}, \hat{g}), (\alpha_T, \beta_T, \tau_T, \sigma_T, f_T, g_T)) = O_p(n^{-s/(2s+1)}).$$

In addition $J(\hat{f}), J(\hat{g}) = O_p(1)$.

Lemma 1 establishes consistency of the PMLE. Furthermore, the estimates of nonparametric covariate effects have the optimal convergence rate $n^{s/(2s+1)}$ (Wahba 1990). Lemma 1 also establishes that $J(\hat{f}), J(\hat{g}) = O_p(1)$, i.e, $\hat{f}$ and $\hat{g}$ have the "right" order of smoothness. The L_2 consistency established in Lemma 1, together with the smoothness and boundedness conditions, can lead to uniform consistency of $\hat{f}$ and $\hat{g}$, i.e., $\sup|\hat{f} - f_T| = o_P(1)$ and $\sup|\hat{g} - g_T| = o_P(1)$. Proof of Lemma 1 is provided in the Appendix. For the estimates of parametric parameters, we have the following results.

Lemma 2. *Under assumptions A1-A4 and additional assumptions provided in the Appendix,*

$$\sqrt{n}\{(\hat{\alpha}, \hat{\beta}, \hat{\tau}, \hat{\sigma}) - (\alpha_T, \beta_T, \tau_T, \sigma_T)\} \rightarrow_D N(0, \Sigma),$$

with the format of Σ specified in the Appendix. Lemma 2 establishes that, despite of the slow convergence rate of $\hat{f}$ and $\hat{g}$, the estimates of parametric parameters are still $\sqrt{n}$ consistent and asymptotically normally distributed. Proof of Lemma 2 is provided in the Appendix.

4. Simulation Study

We conduct simulations to evaluate finite-sample performance of the proposed approach for determination of proportionality and penalized estimation. We generate data from

$$Pr(Y = 0|X, Z) = logit(\boldsymbol{\eta}_1), \quad \text{and for } Y > 0, Y|X, Z = \boldsymbol{\eta}_2 + \boldsymbol{\epsilon}, \tag{13}$$

where $\boldsymbol{\eta}_1 = -4 + 5X_1 - 2.5X_2 + 1.5X_3 + 8sin(6Z_1) + 7Z_2 - 20(Z_2 - 0.5)^2$, $\tau = 0.2$ and $\sigma = 0.5$. We assume the following covariate distributions: $X_1 = 0$ or 1 with probability 1/2; $X_2 = 1, 2, 3$ or 4 with probability 1/4; $X_3 \sim N(0, 1)$; Z_1 is equally spaced between 0

and 1; and $Z_2 \sim Unif[0,1]$. We set the sample size $n = 1000$. We define the "difference function" as $\boldsymbol{\eta}_2 - \tau\boldsymbol{\eta}_1$. Determination of (partial) proportionality then amounts to testing if components of the difference function are equal to zero. As shown in Table 1, ten difference functions are considered. We note that, although some difference functions are linear in Z_i, the corresponding covariate effects in both $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ are still nonlinear. In addition, for a more lucid view, we omit the intercepts in Table 1, which are needed to satisfy the identifiability assumption of $Pf_i = Pg_i = 0$. In the simulation, X_1 is chosen as the anchor.

We first investigate the determination of proportionality. In Table 1, we present power of detecting non-proportionality computed based on 1000 replicates. We can see that, (a) in general, the proposed approach can correctly identify the proportionality structure. More specifically, when proportionality holds for a specific covariate, the power is usually close to zero, indicating a small error rate. When proportionality does not hold, the proposed approach is capable of identifying the non-proportionality with a very high probability. Consider, for example, the last scenario in Table 1 with difference function $0.1X_3 + Z_2$. With probabilities 0.80 and 0.99, the non-proportionality with respect to X_3 and Z_2 can be identified, respectively. The error rates of mistakenly identifying non-proportionality with respect to X_2 and Z_1 are 0.066 and 0.032, respectively; and (b) when the regression coefficients in difference functions (strengths of signals) increase, the power increases. Consider for example scenarios 5 and 6 with difference functions $X_3/3 + 0.5Z_2$ and $X_3/3 + Z_2$, respectively. When the regression coefficient of Z_2 increases from 0.5 to 1, the power increases from 0.40 to 0.95. We have also conducted simulations with other difference functions and/or different sample sizes. Similar satisfactory results are obtained.

For the final models with proportionality properly determined, we also evaluate the penalized estimation results, where the bootstrap inference is based on the procedure described in Section 3.3. We show a representative example of the estimation results in Figure 2, where the

data is generated under simulation scenario 4 with difference function $X_3/3+5Z_1+10Z_1^2+Z_2$. For the covariates with nonparametric effects (Z_1, Z_2), we can see that the mean estimates fit the unknown true functions very well. The 95% confidence intervals provide satisfactory coverage. As expected, the confidence intervals become wider, when it is closer to the boundaries and there are fewer observations. Note that, for identifiability, it has been assumed $Pf_i = Pg_i = 0$. We omit the intercepts (which are needed for the mean zero assumption to be true) in Table 1. The intercepts have been added back in Figure 2, which explains the "shifts" of nonparametric effects and their estimates. We have also examined estimation results for parametric parameters and found negligible biases, satisfactory convergence rates, marginal distributions close to normal, and satisfactory bootstrap coverage. More detailed estimation results are available from the authors. Examination of estimation and inference results with other simulation scenarios leads to similar conclusions.

5. Analysis of MESA Data

The MESA is a population based, multi-center study of subclinical cardiovascular diseases (Bild et al. 2002). The study cohort consists of 6814 subjects with age ranging from 45 to 84 at the baseline. Subjects with missing measurements are removed, which leads to a sample size of 6658 for downstream analysis. We refer to the MESA website http://mesa-nhlbi.org/ for more detailed descriptions of the study design and the cohort.

The distribution of CAC is highly skewed. We consider a simple transformation and analyze $\log(1+CAC)$. As can be seen from Figure 1, $\log(1+CAC)$ has a mixture distribution and the two-part model described in (3) and (4) is thus warranted. In the first part of the two-part model, we assume the commonly used *logit* link function. In the second part of the model, Figure 1 suggests that, it is reasonable to assume a normal distribution for the nonzero $\log(1+CAC)$ values.

Motivated by Han and Kronmal (2006) and McClelland et al. (2006), we consider the

following predictors: gender (female is used as the reference group), race (Caucasian, African-American, Chinese, and Hispanic; Caucasian is used as the reference group), former smoker (binary indicator), current smoker (binary indicator), diabetes (binary indicator), SBP (systolic blood pressure), DBP (diastolic blood pressure), age, BMI (body mass index), LDL cholesterol, and HDL cholesterol. Among the 13 covariates, 7 are binary, which naturally correspond to parametric covariate effects. In addition, published studies and our preliminary analysis suggest linear effects for SBP and DBP. Thus, in the semiparametric models, there are 9 parametric covariate effects and 4 nonparametric ones. Following Han and Kronmal (2006), X_3 is selected as the anchor.

We use the step-wise approach described in Section 3.3 to determine proportionality, and show the results in Table 2. We first fit the model with all covariate effects being proportional. For each covariate effect, we test its proportionality using the bootstrap approach described in Section 3.3. As shown in the first column of Table 2, proportionality of the LDL effect has a p-value < 0.001. Thus we release the proportionality constraint on LDL. At the second step, for each covariate effect other than LDL, we test the proportionality, and present p-values in the second column of Table 2. The HDL effect has a p-value of 0.012. We then fit a model with the proportionality constraints on LDL and HDL released. At the third step, for covariates other than LDL and HDL, we find that releasing the proportionality constraints leads to insignificant p-values. We thus conclude that proportionality holds for all covariates except LDL and HDL.

For the final model with proportionality constraints on all covariates expect LDL and HDL, we present the estimates of parametric regression coefficients in Table 3 and estimates of nonparametric covariate effects in Figure 3. The estimation and bootstrap inference results in Table 3 suggest that the following risk factors are significantly associated with a higher level of CAC: being male, being Caucasian, being a smoker (both former and current), having

diabetes, and having a higher level of SBP. Decrease in DBP is associated with a higher level of CAC, however, the effect is not significant. Those findings are consistent with McClelland et al. (2006) and references therein.

We now examine the estimates of nonparametric covariate effects (Figure 3). For Age and BMI, the covariate effects are proportional. Thus, for the two parts of the models (logistic and linear), the covariate effects take the same shape and only differ by a scale constant. It is interesting that the Age and BMI effects are almost linear, which suggests that it may be possible to further simplify the model by assuming parametric Age and BMI effects. Since the focus of this study is the determination of proportionality, we defer such simplifications to future studies. The bootstrap confidence intervals suggest that both the Age and BMI effects are significant. Increases in Age or BMI are associated with a higher level of CAC, which is consistent with findings in the literature. For LDL and HDL, the proportionality does not hold. The shapes of the covariate effects are significantly different in the two parts of the model. For HDL, its covariate effects have an "U" shape. In the literature, nonparametric modeling of HDL (especially in the context of studying CAC) has not been well investigated. It is very interesting that the HDL effects demonstrate such a shape. Implications of this finding need to be carefully pursued in future biomedical studies. For LDL, it is interesting that the covariate effects are again close to linear. Increase in LDL is associated with a higher probability of nonzero CAC, which is consistent with findings in the literature. The bootstrap confidence intervals suggest significance of the LDL effect. For nonzero CAC values, the LDL effect is negligible.

To provide a more comprehensive understanding of the proposed approach and MESA data, we also fit the full model with no proportionality constraint. Estimation results for the parametric and nonparametric covariate effects are shown in Table 3 and Figure 4, respectively. Comparing estimates under the full and final models, we find that (a) estimates

in the two models are not identical; (b) however, they are reasonably close. This is because estimates under both models are asymptotically consistent; (c) in general, estimates in the final model have smaller variances. In Table 3, all bootstrap standard errors (except for that of X_2 in $\boldsymbol{\eta}_1$) in the full model are larger than or equal to their counterparts in the final model. This is intuitively reasonable, since fewer parameters are estimated in the final model. The improved accuracy has also been observed in Han and Kronmal (2006).

6. Conclusions

In this article, we analyze the CAC in MESA, where it is critical to determine proportionality of covariate effects in semiparametric two-part models. An effective approach, which is composed of penalized maximum likelihood estimation and step-wise determination of proportionality, has been developed. The proposed approach for determination of proportionality advances from published studies by considering more sophisticated models, by allowing for both parametric and nonparametric covariate effects, and by accommodating multiple covariate effects. The proposed penalized estimation approach can accommodate multiple parametric as well as nonparametric covariate effects, and has satisfactory asymptotic properties. Simulation studies and data analysis demonstrate satisfactory finite-sample performance of the proposed approach.

Our analysis of the MESA data suggests that proportionality holds for all covariates except HDL and LDL. Such a finding disproves the hypothesis that the change from a zero to a positive Agaston score and the change from a lower to a higher Agaston score share the same underlying biological process. In contrast, our analysis suggests that the risk factors affect the CAC level via at least two different mechanisms, with the cholesterol having a different mechanism from other risk factors. In addition, our semiparametric analysis suggests that it may be proper to consider linear effects of Age, BMI and LDL in the modeling of CAC. However, the HDL effect needs to be described in a nonparametric way.

Since the reduction from nonlinear to linear covariate effects is not the focus of our study, we defer such investigations to future studies. "Directions" of covariate effects (i.e., whether they are positively or negatively associated with CAC) are consistent with the literature, which further confirms published findings.

For a more lucid description of the methodology, we have made the simplified assumption of additive nonparametric covariate effects. More general assumptions that allow "interactions" among covariates can be assumed. We note that such assumptions may dramatically increase the computational cost and will not be pursued. Motivated by previous studies, we have assumed smooth covariate effects. It is possible to replace the smoothness assumption with other (e.g. monotone) assumptions. Determination of proportionality requires selecting the anchor covariate. As a rule of thumb, we propose fitting marginal models with only one covariate at a time, and selecting the covariate with the smallest marginal p-value. In our data analysis, we select the same anchor as Han and Kronmal (2006). In this article, a forward step-wise approach has been adopted. In other studies that involve selection of covariate effects, it has been suggested that there are approaches more effective than the forward step-wise approach. Our simulation studies suggest satisfactory performance of the simple forward step-wise approach, although we note that it can be potentially improved.

Acknowledgements

This study has been supported by N01-HC95159 from NHLBI (Kronmal) and DMS 0805984 from NSF (Zhou and Ma). We thank the investigators, the staff, and the participants of MESA for their valuable contributions. A full list of participating MESA investigators and institutions can be found at http://www.mesa-nhlbi.org.

References

Albert, P.S., Follmann, D.A, and Barnhart, H.X. (1997). A generalized estimating equation approach for modeling random length binary vector data. *Biometrics*, **53**, 1116-1124.

Bild, D.E., Bluemke, D.A., Burke, G.L., Detrano, R, Diez-Roux, A.V., Folsom, A.R., Greenland, P., Jacob, D.R. Jr, Kronmal, R., Liu, K., Nelson, J.C., O'Leary, D., Saad, M.F., Shea, S., Szklo, M. and Tracy, R.P. (2002). Multi-ethnic study of atherosclerosis: objectives and design. *American Journal of Epidemiology*, **156**, 871–881.

Clarkson, D.B. and Zhan, Y. (2002). Using spherical-radial quadrature to fit generalized linear mixed effects models. *Journal of Computational and Graphical Statistics*, **11**, 639–659.

Cragg, J.G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica*, **39**, 829–844.

Crainiceanu, C. and Ruppert, D. and Claeskens, G. and Wand, M. (2005). Exact likelihood ratio tests for penalised splines. *Biometrika*, **92**, 91-103.

Guo, W. (2002). Inference in smoothing spline analysis of variance. *Journal of The Royal Statistical Society, Ser B*, **64**, 887-898.

Han, C. and Kronmal, R.A. (2006). Two-part models for analysis of Agatston scores with possible proportionality constraints. *Communications in Statistics–Theory and Methods*, **35**, 99-111.

Lam, K.F. and Xue, H. (2005). A semiparametric regression cure model with current status data. *Biometrika*, **92**, 573–586.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**, 1–14.

Liu, A., Meiring, W. and Wang, Y. (2005). Testing generalized linear models using

smoothing spline methods. *Statistica Sinica*, **15**, 235-256.

Ma, S. (2009). Cure model with current status data. *Statistica Sinica*, **19**, 233-249.

Ma, S. and Kosorok, M.R. (2005). Robust semiparametric M-estimation and the weighted bootstrap. *Journal of Multivariate Analysis*, **96**, 190-217.

McClelland, R.L., Chung, H., Detrano, R., Post, W., and Kronmal, R.A. (2006). Distribution of coronary artery calcium by race, gender, and age. Results from the Multi-Ethnic Study of Atherosclerosis (MESA). *Ciuculation*, **113**, 30-37.

Monahan, J. and Genz, A. (1997). Spherical-radial integration rules for a Bayesian computation. *Journal of the American Statistical Association*, **92**, 664-674.

Moulton, L.H., Curriero, F.C. and Barroso, P.F. (2002) Mixture models for quantitative HIV RNA data. *Statistical Methods in Medical Research*, **11**, 317-325.

van de Geer, S. (2000). *Empirical Processes in M-Estimation*. Cambridge Series in Statistical and Probabilistic Mathematics.

Wahba, G. (1990). *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathematics, SIAM.

Table 1

Simulation study: power of testing non-proportionality with various difference functions.

	Power			
Difference function	X_2	X_3	Z_1	Z_2
$X_3/3$	0.045	1	0.021	0.054
$5Z_1 + Z_1^2 + 0.9Z_2$	0.051	0.064	0.635	0.806
$0.8X_2 + 5Z_1 + 2Z_1^2 + 0.8Z_2$	0.900	0.046	0.820	0.620
$X_3/3 + 5Z_1 + 10Z_1^2 + Z_2$	0.076	0.980	1	0.920
$X_3/3 + 0.5Z_2$	0.062	1	0.033	0.400
$X_3/3 + Z_2$	0.079	1	0.048	0.950
$0.3X_2 + X_3/3$	0.220	1	0.042	0.051
$0.5X_2 + X_3/3$	0.560	1	0.035	0.050
$0.05X_3 + Z_2$	0.045	0.160	0.038	0.960
$0.1X_3 + Z_2$	0.066	0.800	0.032	0.990

Table 2

Analysis of the CAC in MESA. Each column shows the p-values of proportionality for the corresponding terms in the model at each step of the forward step-wise relaxation of proportionality constraint. X_3 has been selected as the anchor.

Relaxed predictor	P-value		
Gender: Male (X_1)	0.800	0.607	0.610
Race: Chinese (X_2)	0.014	0.214	0.233
Race: African-American (X_3)	–	–	–
Race: Hispanic (X_4)	0.770	0.286	0.267
Former smoker (X_5)	0.713	0.339	0.311
Current smoker (X_6)	0.621	0.964	0.784
Diabetes (X_7)	0.140	0.393	0.307
SBP (X_8)	0.547	0.256	0.671
DBP (X_9)	0.374	0.387	0.233
Age (Z_1)	0.104	0.607	0.285
BMI (Z_2)	0.612	0.440	0.767
LDL (Z_3)	**<0.001**		
HDL (Z_4)	0.025	**0.012**	

Table 3

Analysis of the CAC in MESA. Parametric regression coefficients in the full model (with no proportionality constraint), and the final model (with proportionality properly determined). Estimates (bootstrap standard errors) in the logistic ($\boldsymbol{\eta}_1$) and linear ($\boldsymbol{\eta}_2$) models.

	Full model		Final model	
Predictor	$\boldsymbol{\eta}_1$	$\boldsymbol{\eta}_2$	$\boldsymbol{\eta}_1$	$\boldsymbol{\eta}_2$
Gender: Male (X_1)	0.945 (0.092)	0.618 (0.099)	0.960 (0.078)	0.651 (0.053)
Race: Chinese (X_2)	-0.119 (0.070)	-0.285 (0.081)	-0.211 (0.078)	-0.143 (0.053)
Race: African-American (X_3)	-0.787 (0.071)	-0.398 (0.085)	-0.727 (0.063)	-0.493 (0.047)
Race: Hispanic (X_4)	-0.628 (0.074)	-0.358 (0.073)	-0.594 (0.063)	-0.402 (0.045)
Former smoker (X_5)	0.370 (0.072)	0.213 (0.071)	0.354 (0.052)	0.240 (0.036)
Current smoker (X_6)	0.609 (0.094)	0.328 (0.096)	0.573 (0.078)	0.388 (0.052)
Diabetes (X_7)	0.243 (0.070)	0.275 (0.068)	0.299 (0.055)	0.203 (0.038)
SBP (X_8)	0.009 (0.002)	0.004 (0.002)	0.008 (0.002)	0.005 (0.001)
DBP (X_9)	-0.0034 (0.004)	0.0032 (0.004)	-0.0009 (0.004)	-0.0006 (0.002)
τ			0.678 (0.037)	
σ	1.677 (0.021)		1.680 (0.021)	

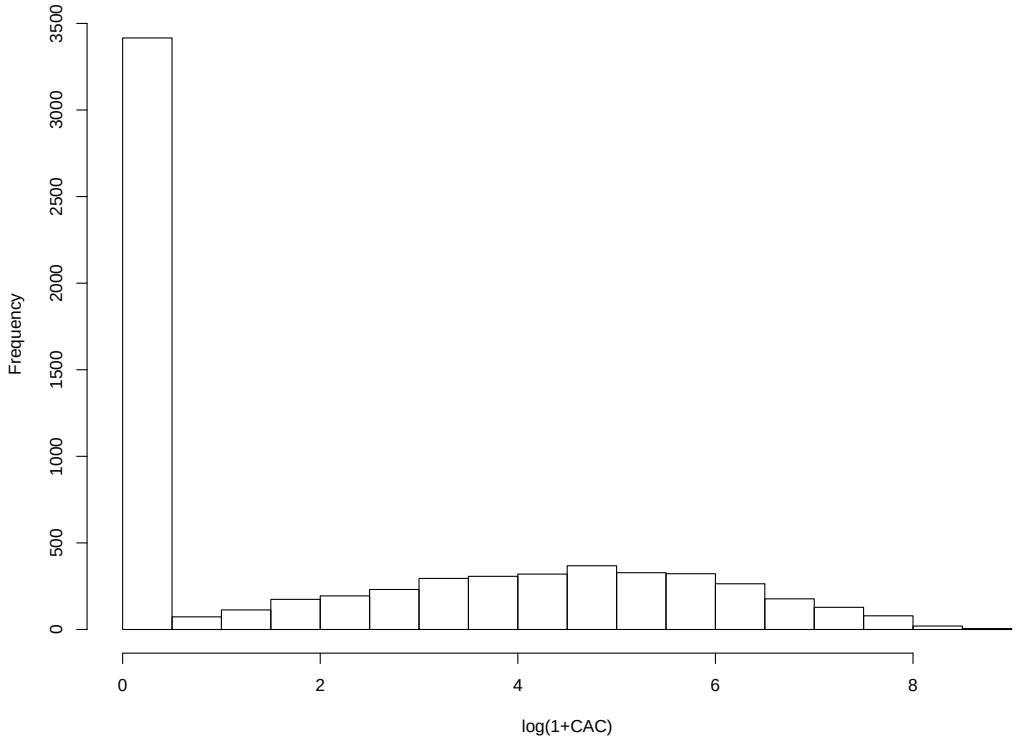

Figure 1. MESA data: Histogram of $\log(1 + CAC)$.

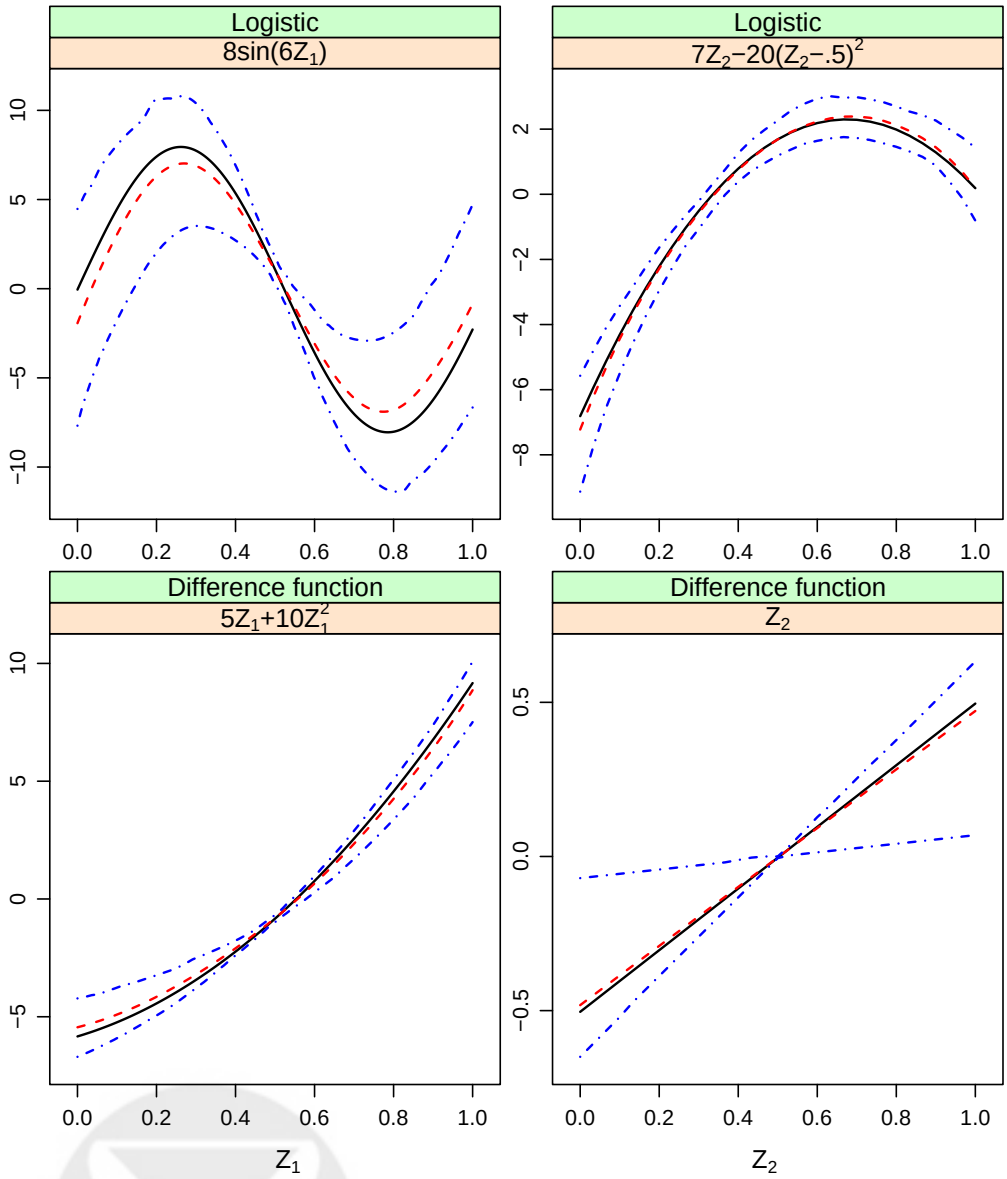

Figure 2. Simulation study with difference function $X_3/3+5Z_1+10Z_1^2+Z_2$: estimation and inference results for nonparametric covariate effects. Solid black line: true covariate effect; Red dashed line: mean estimates; Blue dash-dotted lines: mean 95% confidence intervals.

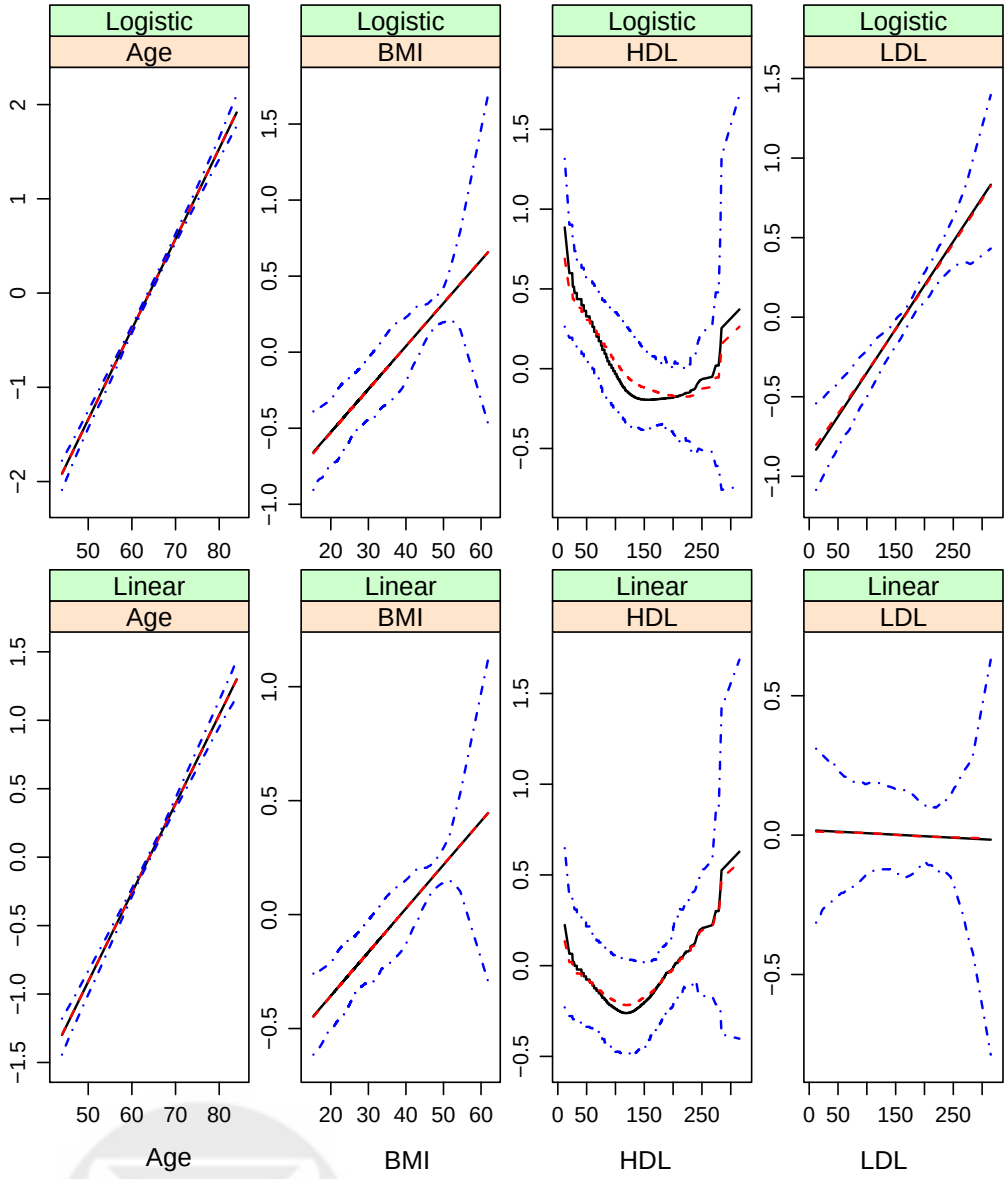

Figure 3. Analysis of the CAC in MESA, **final model** with proportionality properly determined. Estimated nonparametric covariate effects in both the logistic and linear parts of the model. Solid black line: estimate; Red dashed line: mean estimate from bootstrap samples; Blue dash-dotted lines: 95% confidence intervals.

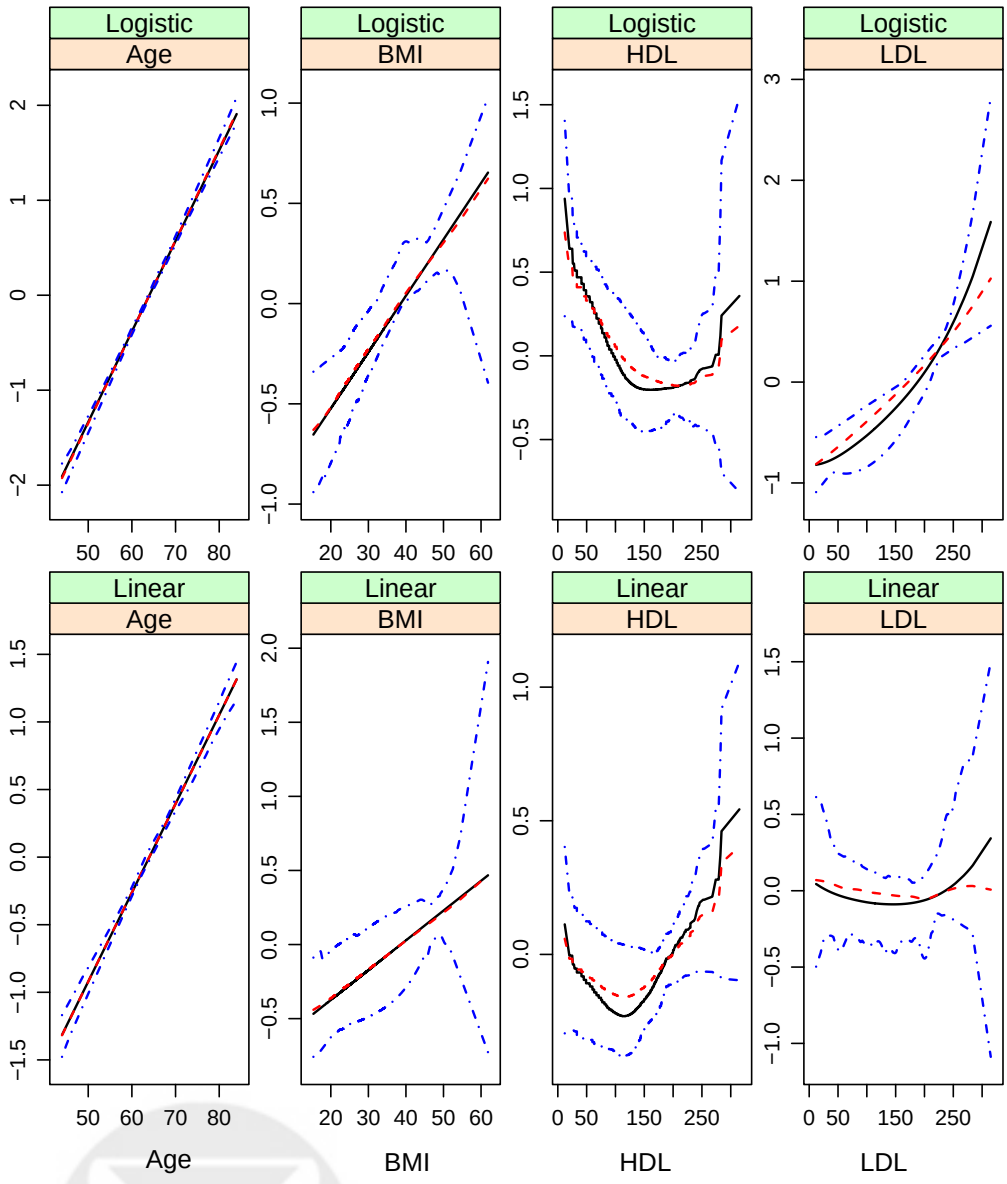

Figure 4. Analysis of the CAC in MESA, **full model** with no proportionality constraint. Estimated nonparametric covariate effects in both the logistic and linear parts of the model. Solid black line: estimate; Red dashed line: mean estimate from bootstrap samples; Blue dash-dotted lines: 95% confidence intervals.

Appendix

Proof of Lemma 1.

DEFINITION (Bracketing number). Let $(\mathbb{F}, ||\cdot||)$ be a subset of a normed space of real function h on some set. Given two functions h_1 and h_2, the bracket $[h_1, h_2]$ is the set of all functions h with $h_1 \leq h \leq h_2$. An ϵ bracket is a bracket $[h_1, h_2]$ with $||h_1 - h_2|| \leq \epsilon$. The bracketing number $N_{[]}(\epsilon, \mathbb{F}, ||\cdot||)$ is the minimum number of ϵ brackets needed to cover $\mathbb{F}$. The entropy with bracketing is the logarithm of the bracketing number.

van de Geer (2002) proves that for the functional class

$$\tilde{\mathbb{H}} = \{h : [0,1] \to [0,1] \int (h^{(s)}(x))^2 dx < 1\},$$

$\log N_{[]}(\epsilon, \tilde{\mathbb{H}}, L_2(P)) \leq K_2 \epsilon^{-1/s}$, for a fixed constant K_2, $s \geq 1$, and all ϵ.

Under the boundedness assumptions A1 and A2 and the differentiability of the log-likelihood function, we have

$$\log N_{[]}(\epsilon, l(\alpha, \beta, \tau, \sigma, f, g), L_2(P)) \leq K_3 \epsilon^{-1/s}, \tag{14}$$

for a fixed constant K_3.

Examination of the log-likelihood function suggests that if $\hat{\alpha}, \hat{\beta}, \hat{\tau}, \hat{\sigma} \to \infty$, then $P_n l \to -\infty$. Thus, we are able to focus on the set of bounded $\hat{\alpha}, \hat{\beta}, \hat{\tau}, \hat{\sigma}$, although the actual bound remains unknown. In addition, following Wahba (1990), it can be shown that under assumption A2, $\hat{f}$ and $\hat{g}$ are spline functions. Specifically, suppose $\tilde{f}$ and $\tilde{g}$ maximize the penalized log-likelihood function. Then there exist spline functions $\hat{f}$ and $\hat{g}$, such that $\hat{f}(Z) = \tilde{f}(Z)$ and $\hat{g}(Z) = \tilde{g}(Z)$ at all the observed Z values and $J(\hat{f}) \leq J(\tilde{f})$ and $J(\hat{g}) \leq J(\tilde{g})$.

From the definition of the PMLE, we have

$$P_n l(\hat{\alpha}, \hat{\beta}, \hat{\tau}, \hat{\sigma}, \hat{f}, \hat{g}) - \lambda_f^2 J^2(\hat{f}) - \lambda_g^2 J^2(\hat{g}) \geq P_n l(\alpha_T, \beta_T, \tau_T, \sigma_T, f_T, g_T) - \lambda_f^2 J^2(f_T) - \lambda_g^2 J^2(g_T).$$

From the properties of the likelihood function, we have

$$Pl(\hat{\alpha}, \hat{\beta}, \hat{\tau}, \hat{\sigma}, \hat{f}, \hat{g}) \leq Pl(\alpha_T, \beta_T, \tau_T, \sigma_T, f_T, g_T).$$

Combining the above two equations, we get

$$\begin{aligned}
&\lambda_f^2 J^2(\hat{f}) + \lambda_g^2 J^2(\hat{g}) + P(l(\alpha_T, \beta_T, \tau_T, \sigma_T, f_T, g_T) - l(\hat{\alpha}, \hat{\beta}, \hat{\tau}, \hat{\sigma}, \hat{f}, \hat{g})) \\
&\leq \lambda_f^2 J^2(f_T) + \lambda_g^2 J^2(g_T) + (P_n - P)(l(\hat{\alpha}, \hat{\beta}, \hat{\tau}, \hat{\sigma}, \hat{f}, \hat{g}) - l(\alpha_T, \beta_T, \tau_T, \sigma_T, f_T, g_T)). \quad (15)
\end{aligned}$$

In addition, the entropy result in (14) implies that

$$\begin{aligned}
&(P_n - P)(l(\alpha_T, \beta_T, \tau_T, \sigma_T, f_T, g_T) - l(\hat{\alpha}, \hat{\beta}, \hat{\tau}, \hat{\sigma}, \hat{f}, \hat{g})) \\
&\quad = o_P(n^{-1/2})(1 + J(f_T) + J(g_T) + J(\hat{f}) + J(\hat{g})). \quad (16)
\end{aligned}$$

Combining equations (15) and (16) with assumption A4, we have

$$\lambda_f J(\hat{f}) = o_P(1) \text{ and } \lambda_g J(\hat{g}) = o_P(1). \quad (17)$$

Under assumption A3, equations (15) and (16) imply that

$$\begin{aligned}
&K_1 d^2((\alpha_T, \beta_T, \tau_T, \sigma_T, f_T, g_T), (\hat{\alpha}, \hat{\beta}, \hat{\tau}, \hat{\sigma}, \hat{f}, \hat{g})) \\
&\quad \leq o_P(1) + o_P(n^{-1/2})(1 + J(f_T) + J(g_T) + J(\hat{f}) + J(\hat{g})).
\end{aligned}$$

This equation and equation (17) lead to consistency of the PMLE. To prove the rate of convergence, we use the following result.

(Theorem in van de Geer 2000). Consider a uniformly bounded class of functions Γ, with $\sup_{\gamma \in \Gamma} |\gamma - \gamma_0|_\infty < \infty$ and a fixed $\gamma_0 \in \Gamma$, and $\log N_{[]}(\epsilon, \Gamma, P) \leq K_4 \epsilon^{-b}$ for all $\epsilon > 0$, where $b \in (0, 2)$ and K_4 is a fixed constant. Then for $\delta_n = n^{-1/(2+b)}$,

$$\sup_{\gamma \in \Gamma} \frac{|(P_n - P)(\gamma - \gamma_0)|}{||\gamma - \gamma_0||_2^{1-b/2} \vee \sqrt{n}\delta_n^2} = O_p(n^{-1/2}), \quad (18)$$

where $x \vee y = max(x, y)$.

Under the compactness assumptions A1 and A2 and considering the differentiability of the log-likelihood function, we have

$$\begin{aligned} K_1 d^2((\hat{\alpha}, \hat{\beta}, \hat{\tau}, \hat{\sigma}, \hat{f}, \hat{g}), (\alpha_T, \beta_T, \tau_T, \sigma_T, f_T, g_T)) &\leq P(l(\alpha_T, \beta_T, \tau_T, \sigma_T, f_T, g_T) - l(\hat{\alpha}, \hat{\beta}, \hat{\tau}, \hat{\sigma}, \hat{f}, \hat{g})) \\ &\leq K_5 d^2((\hat{\alpha}, \hat{\beta}, \hat{\tau}, \hat{\sigma}, \hat{f}, \hat{g}), (\alpha_T, \beta_T, \tau_T, \sigma_T, f_T, g_T)), \end{aligned} \tag{19}$$

where K_5 is a fixed constant. Combining (18) with $b = \frac{1}{s}$, equations (19), and (15), we have

$$\begin{aligned} &\lambda_f^2 J^2(\hat{f}) + \lambda_g^2 J^2(\hat{g}) + K_1 d^2((\hat{\alpha}, \hat{\beta}, \hat{\tau}, \hat{\sigma}, \hat{f}, \hat{g}), (\alpha_T, \beta_T, \tau_T, \sigma_T, f_T, g_T)) \\ &\leq \lambda_f^2 J^2(f_T) + \lambda_g^2 J^2(g_T) + O_P(n^{-1/2})(1 + J(f_T) + J(\hat{f}) + J(g_T) + J(\hat{g})) \\ &\quad \times \{d^{1-1/2s}((\hat{\alpha}, \hat{\beta}, \hat{\tau}, \hat{\sigma}, \hat{f}, \hat{g}), (\alpha_T, \beta_T, \tau_T, \sigma_T, f_T, g_T)) \vee n^{\frac{1-2s}{2(2s+1)}}\}. \end{aligned} \tag{20}$$

Note that all the three terms on the left hand side are positive. Compare each term with the right hand side. Simple calculations give that

$$J(\hat{f}) = O_P(1) \quad \text{and} \quad J(\hat{g}) = O_P(1),$$

$$d((\hat{\alpha}, \hat{\beta}, \hat{\tau}, \hat{\sigma}, \hat{f}, \hat{g}), (\alpha_T, \beta_T, \tau_T, \sigma_T, f_T, g_T)) = O_P(n^{-s/(2s+1)}).$$

Proof of Lemma 2.

To prove the $\sqrt{n}$ consistency and asymptotic normality result in Lemma 2, we apply Theorem 1 in Ma and Kosorok (2005). Application of this theorem requires the following conditions to hold: (a) consistency and rate of convergence, which has been established in Lemma 1; (b) finite asymptotic variance, which is shown below; (c) stochastic equicontinuity, which can be established using the entropy result and the consistency result; and (d) smoothness of the model, which holds given the differentiability of the likelihood function.

Thus, to prove Lemma 2, we only need to establish the non-singularity of the information matrix. Denote $\dot{l}_\alpha, \dot{l}_\beta, \dot{l}_\tau, \dot{l}_\sigma$ as the partial derivative of the log-likelihood function with respect to $\alpha, \beta, \tau, \sigma$. For $t_f, t_g \sim 0$, consider $f_t = f + t_f\xi_f$ and $g_t = g + t_g\xi_g$, such that f_t, g_t still satisfy assumption A2. Denote the space generated by $\xi_f \otimes \xi_g$ as $\mathbb{B}$. The score operators for f and g are $\dot{l}_f[\xi_f] = \lim_{t_f \to 0} \frac{l(\alpha,\beta,\tau,\sigma,f_t,g) - l(\alpha,\beta,\tau,\sigma,f,g)}{t_f}$ and $\dot{l}_g[\xi_g] = \lim_{t_g \to 0} \frac{l(\alpha,\beta,\tau,\sigma,f,g_t) - l(\alpha,\beta,\tau,\sigma,f,g)}{t_g}$. Denote $\dot{l}_1 = (\dot{l}_\alpha, \dot{l}_\beta, \dot{l}_\tau, \dot{l}_\sigma)'$ as the score function for the parametric parameters and $\dot{l}_{f,g}[\xi_f, \xi_g] = (\dot{l}_f[\xi_f], \dot{l}_g[\xi_g])$ as the score operator for the nonparametric parameters.

Project $\dot{l}_1$ onto the space generated by $\dot{l}_{f,g}[\xi_f, \xi_g] = (\dot{l}_f[\xi_f], \dot{l}_g[\xi_g])$. The efficient score for (α, β, τ) is $U = \dot{l}_1 - \dot{l}_{f,g}\left[\frac{P(\dot{l}_1)\dot{l}_{f,g}|Z}{P(\dot{l}^{\otimes}_{f,g}|Z)}\right]$. We further assume

(A5). $P(U'U)$ is component-wise bounded and positive definite.

Then $P(U'U)$ is the information matrix, and $\Sigma = P^{-1}(U'U)$ is the asymptotic variance matrix.