Cure Rate Models with Nonparametric Form of Covariate Effects

Tianlei Chen

Dissertation submitted to the Faculty of the

Virginia Polytechnic Institute and State University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Statistics

Pang Du, Chair

Yili Hong

Inyong Kim

George R. Terrell

April 27, 2015

Blacksburg, Virginia

Keywords: Nonparametric Function Estimation, Smoothing Spline, Penalized Likelihood Method, Survival Analysis, Cure Rate Model

Cure Rate Models with Nonparametric Form of Covariate Effects

Tianlei Chen

(ABSTRACT)

This thesis focuses on development of spline-based hazard estimation models for cure rate data. Such data can be found in survival studies with long term survivors. Consequently, the population consists of the susceptible and non-susceptible subpopulations with the latter termed as "cured". The modeling of both the cure probability and the hazard function of the susceptible subpopulation is of practical interest. Here we propose two smoothing-splines-based models falling respectively into the popular classes of two component mixture cure rate models and promotion time cure rate models.

Under the framework of two component mixture cure rate model, Wang, Du and Liang (2012) have developed a nonparametric model where the covariate effects on both the cure probability and the hazard component are estimated by smoothing splines. Our first development falls under the same framework but estimates the hazard component based on the accelerated failure time model, instead of the proportional hazards model in Wang, Du and Liang (2012). Our new model has better interpretation in practice.

The promotion time cure rate model, motivated from a simplified biological interpretation of cancer metastasis, was first proposed only a few decades ago. Nonetheless, it has quickly become a competitor to the mixture models. Our second development aims to provide

a nonparametric alternative to the existing parametric or semiparametric promotion time models.

ACKNOWLEDGMENTS

Foremost, I would like to express my sincere gratitude to my advisor Prof. Pang Du for the continuous support of my Ph.D study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D study.

Besides my advisor, I would like to thank the rest of my thesis committee: Prof. Yili Hong, Prof. Inyong Kim, and Prof. George R. Terrell, for their encouragement, insightful comments, and hard questions.

Furthermore, I want to thank all the faculty and staff members of the Department of Statistics for their help during my five years at Virginia Tech. I want to thank all my fellow students who have offered assistance, encouragement and friendship during the course of my study.

Last but not the least, I would like to thank my family: my parents, for giving birth to me at the first place and supporting me spiritually throughout my life.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Figure Page

1. INTRODUCTION

Due to rapid developments and significant progress in medical and health sciences, we often see medical studies with long term survivors. This calls for survival models incorporating a cure rate (cure rate models) to account for the subjects not at risk of death (or relapse).

One category of cure rate model is two-component mixture cure model first proposed by Berkson and Gage [1952]. It is dominant in cure rate studies since it was proposed. The model assumes that the study population is a mixture of a susceptible subpopulation who are destined to experience the event if observed long enough and a nonsusceptible subpopulation who are considered free of the event or as "cured". The statistical model for such data often consists of one component for the cure probability and the other component for the hazard function of susceptible subjects. In Farewell [1982], the cure rate component was modeled as parametric logistic regression model and the survival component assumed Weibull distribution. Kuk and Chen [1992] extended Farewell [1982] by formulating the survival component with semiparametric Cox proportional hazards model. They applied a marginal likelihood approach and used an estimation method involving Monte Carlo simulation. In Peng and Dear [2000] and Sy and Taylor [2000], the model is similar in spirit to that of Kuk and Chen [1992], but the estimation is through an EM algorithm. Lu and Ying [2004] used the mixture formulation to extend a class of semiparametric transformation models proposed

by Cheng et al. [1995] to incorporate cure fractions. In Othus et al. [2009], the semiparametric transformation model that allows for covariates as well as dependent censoring was proposed. Recently, Wang et al. [2012] proposed a two-component mixture cure rate model with nonparametric forms for both the cure probability and the hazard rate function. More detailed literature review for cure rate data is in Section 2.1. In this dissertation, we propose a nonparametric two-component mixture model where hazard function is estimated by smoothing splines under the framework of accelerated failure time model.

Although the two component mixture model is a popular approach, it has some drawbacks from both a Bayesian and frequentist perspective, as pointed out by Chen et al. [1999] and Ibrahim et al. [2001]. An alternative but equally general model, namely promotion time cure rate model has been proposed in which the survival distribution for cured subjects and non-cured subjects are integrated into one improper survival time distribution, see, e.g., Yakovlev and Tsodikov [1996], Tsodikov [1998], Chen et al. [1999], Tsodikov et al. [2003], Zeng et al. [2006]. This type of model first appeared in Yakovlev and Tsodikov [1996] and Tsodikov [1998] who also noted that the model provided a natural way to extend the proportional hazards regression model. Chen et al. [1999] gave a biological interpretation of the model and proposed a Bayesian analysis method. They also pointed out a mathematical connection between the promotion cure model and the two-component mixture cure models. Zeng et al. [2006] considered a more general form of model with a parameter that offers more transformation. This type of model has some distinct advantages such as providing a biologically meaningful interpretation of the model result and allowing construction of a

rich class of nonlinear transformation regression models to describe complex covariate effects. More detailed literature review for the promotion time cure rate model is in Section 3.1. The existing promotion time cure rate models have a common limitation in that they all model covariate effects in a parametric form whose validity is generally not justified in practice. The strict parametric assumption can be particularly problematic at the exploratory stage of a study. This calls for more flexible nonparametric modeling of covariate effects. To avoid the model misspecification in a parametric analysis, we propose a nonparametric promotion time cure rate model where the link function for covariates are estimated with smoothing splines through penalized likelihood method.

The framework of penalized likelihood method adopted in this study is based on Wahba [1990] and Gu [2004]. Given stochastic data "generated" according to an unknown "pattern" function η_0, the penalized likelihood method estimates η_0 by minimizing a score of the form

$$L(\eta|\text{data}) + \frac{\lambda}{2}J(\eta), \tag{1.1}$$

where $L(\eta)$, usually the negative log likelihood, measures the goodness-of-fit of η, $J(\eta)$, the roughness penalty, measures the smoothness of η, and the smoothing parameter $\lambda(> 0)$ controls the trade off. The minimization of (1.1) is done in a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ of functions. RKHS provides a theoretical basis for Smoothing Spline ANOVA (SSANOVA) model and unified framework for modeling various data. For multivariate η,

it can be decomposed into main effects and interactions similar to the classical ANOVA decomposition.

Through proper specifications of η and $J(\eta)$ in a variety of problem settings, (1.1) yields nonparametric models for Gaussian and non-Gaussian regression, probability density estimation, hazard rate estimation, etc. Kimeldorf and Wahba [1970a], Kimeldorf and Wahba [1970b] and Kimeldorf and Wahba [1971] first proposed penalized least squares regression in univariate case. The general problem of penalized least squares regression with multiple penalty terms was formulated by Wahba [1986]. Non-Gaussian regression in such context can be found in Gu [1990] and Gu and Xiang [2001]. The penalized likelihood method in the context of density estimation was studied by Good and Gaskins [1971], Wahba et al. [2001] and Gu and Qiu [1993]. The formulation of penalized likelihood hazard estimation used in this dissertation was proposed by Gu [1996]. The setting relevant to our cure rate data problem in Chapter 2 is regression with responses from exponential family, so in this chapter we introduce the smoothing spline estimation details for this setting.

The function estimates in our cure rate problem in Chapter 2 are computed through an EM algorithm. Interval estimates are constructed through an extension of the well known Louis formula for EM estimation. Hence at the end of this chapter, we give a brief introduction of the EM algorithm, see Dempster et al. [1977] and Louis' formula, see Louis [1982].

1.1 Penalized Likelihood for Data from Exponential Family

In general, consider an exponential family distribution with density of the form

$$f(y|x) = \exp\{(y\eta(x) - b(\eta(x)))/a(\phi) + c(y,\phi)\} \tag{1.2}$$

where $a > 0$, b, and c are known functions, η is the canonical parameter of interest dependent on a covariate x, and ϕ is either known or considered as a nuisance parameter that is independent of x. The formulation (1.2) includes Gaussian, Binomial, Poisson distributions as special cases. The penalized likelihood for estimating η is

$$-\frac{1}{n}\sum_{i=1}^{n}\{Y_i\eta(x_i) - b(\eta(x_i))\} + \frac{\lambda}{2}J(\eta). \tag{1.3}$$

Note that in (1.3), the dispersion function $a(\phi)$ is absorbed by the smoothing parameter and the term $c(y,\phi)$ is dropped since it is not relevant to the optimization with respect to the canonical parameter η. Examples of penalized likelihood estimation for non-Gaussian data can be found in Gu [1990] and Wahba et al. [1995], among others. Particularly, for binary data with observed binary outcomes y_i and covariates x_i, the penalized likelihood is

$$\sum_{i=1}^{n}\left\{y_i\eta(x_i) - \log[1 + e^{\eta(x_i)}]\right\}, \tag{1.4}$$

where $\eta(x_i) = log(p_i/(1-p_i))$ is the logit function of the probability $p_i = P(Y_i = 1|x_i)$.

1.1.1 Model Construction

Note that η in (1.4) is a restriction free function. Hence we don't need any other constraints on η besides the smoothness requirement imposed by the penalty $J(\cdot)$. Next, we give brief descriptions of RKHS configurations under different covariate settings for data from exponential family.

Example 1.1.1 (One covariate $\mathcal{V}$) *One only has the domain $\mathcal{V} = [0, 1]$. A choice of $J(\eta)$ is $\int_0^1 (\eta'')^2 dv$, which yields the popular cubic splines. A choice of $\tilde{J}(f, g)$ is $(\int_0^1 f\, dv)(\int_0^1 g\, dv) + (\int_0^1 \dot{f}\, dv)(\int_0^1 \dot{g}\, dv)$, yielding $\mathcal{H}_J = \{\eta : \int_0^1 \eta dv = \int_0^1 \dot{\eta} dv = 0, J(\eta) < \infty\}$ and the RK $R_J(v_1, v_2) = k_2(v_1)k_2(v_2) - k_4(v_1 - v_2)$, where $k_\nu = B_\nu/\nu!$ are scaled Bernoulli polynomials. The null space $\mathcal{N}_J$ has a basis $\{1, k_1(v)\}$, where $k_1(v) = v - 0.5$. See, e.g., Section 2.3.3 in Gu [2002].* □

Example 1.1.2 (One continuous and one categorical covariates $\mathcal{V} \times \mathcal{U}$) *One continuous covariate, say $\mathcal{V} = [0, 1]$, could be any continuous variable and one categorical covariates, say $\mathcal{U} = \{1, 2\}$, could be a simple categorical covariate representing control and treatment. Functions on $\mathcal{U}$ are actually vectors in R^2. Taking $J_{\langle u \rangle}(\eta) = [\eta(1) - \eta(2)]^2/2$ and $\tilde{J}_{\langle u \rangle}(\eta) = [\eta(1) + \eta(2)]^2/2$, the RKHS $\mathcal{H}_{\langle u \rangle} = R^2$ on the covariate domain can be decomposed as*

$$\mathcal{H}_{\langle u \rangle} = \mathcal{H}_{0\langle u \rangle} \oplus \mathcal{H}_{1\langle u \rangle} = \{\eta : \eta(1) = \eta(2)\} \oplus \{\eta : \eta(1) + \eta(2) = 0\}$$

with RKs $R_{0\langle u\rangle}(u_1,u_2)=1/2$, $R_{1\langle u\rangle}(u_1,u_2)=I_{[u_1=u_2]}-1/2$. *On the other hand, the construction in Example 1.1.1 gives a decomposition of the RKHS* $\mathcal{H}_{\langle v\rangle}$ *on the* $\mathcal{V}$ *domain*

$$\begin{aligned}\mathcal{H}_{\langle v\rangle} &= \left\{\eta : \textstyle\int_0^1 (\eta'')^2 dx < \infty\right\} = \mathcal{H}_{00\langle v\rangle} \oplus \mathcal{H}_{01\langle v\rangle} \oplus \mathcal{H}_{1\langle v\rangle} \\ &= span\{1\} \oplus span\{k_1(x)\} \oplus \left\{\eta : \textstyle\int_0^1 \eta dx = \int_0^1 \dot{\eta} dx = 0,\ \int_0^1 (\eta'')^2 dx < \infty\right\},\end{aligned}$$

with RKs $R_{00\langle v\rangle}(v_1,v_2)=1$, $R_{01\langle v\rangle}(v_1,v_2)=k_1(v_1)k_1(v_2)$, *and* $R_{1\langle v\rangle}(v_1,v_2)=k_2(v_1)k_2(v_2)-k_4(v_1-v_2)$. *Taking tensor product of* $\mathcal{H}_{\langle v\rangle}$ *and* $\mathcal{H}_{\langle u\rangle}$, *one obtains six tensor sum terms* $\mathcal{H}_{\nu,\mu}=\mathcal{H}_{\nu\langle v\rangle}\otimes\mathcal{H}_{\mu\langle u\rangle}$ *on* $\mathcal{V}\times\mathcal{U}$, $\nu=00,01,1$ *and* $\mu=0,1$, *with RKs* $R_{\nu,\mu}(x_1,x_2)=R_\nu(v_1,v_2)R_\mu(u_1,u_2)$, *where* $x_i=(v_i,u_i)$. *The two subspaces with* $\nu=00,01$ *and* $\mu=0$ *are of one-dimension each, and can be lumped together as* $\mathcal{N}_J$. *The other four subspaces can be put together as* $\mathcal{H}_J$ *with the RK*

$$R_J=\theta_{00,1}R_{00\langle v\rangle,1\langle u\rangle}+\theta_{01,1}R_{01\langle v\rangle,1\langle u\rangle}+\theta_{1,0}R_{1\langle v\rangle,0\langle u\rangle}+\theta_{1,1}R_{1\langle v\rangle,1\langle u\rangle},$$

where $\theta_{\nu,\mu}$ *are a set of extra smoothing parameters adjusting the relative weights of the roughness of different components. For more detail about multiple term RKHS with multiple smoothing parameters, see Section 2.4.5 in Gu [2002].*

For interpretation, the six subspaces readily define an ANOVA decomposition

$$\eta(v,u)=\eta_\emptyset+\eta_v(v)+\eta_u(u)+\eta_{v,u}(v,u)$$

for functions on $\mathcal{V} \times \mathcal{U}$, with $\eta_\emptyset \in \mathcal{H}_{00\langle v\rangle} \otimes \mathcal{H}_{0\langle u\rangle}$ being the constant term, $\eta_v \in \{\mathcal{H}_{01\langle v\rangle} \oplus \mathcal{H}_{1\langle v\rangle}\} \otimes \mathcal{H}_{0\langle u\rangle}$ the v main effect, $\eta_u \in \mathcal{H}_{00\langle v\rangle} \otimes \mathcal{H}_{1\langle u\rangle}$ the u main effect, and $\eta_{v,u} \in \{\mathcal{H}_{01\langle v\rangle} \oplus \mathcal{H}_{1\langle v\rangle}\} \otimes \mathcal{H}_{1\langle u\rangle}$ the interaction. This example can also be generalized to the case of a multi-level categorical variable or an ordinal variable, with the configuration for the latter slightly different from the one presented here. See, e.g., Section 2.2 in Gu [2002].

Example 1.1.3 (Two continuous covariates $\mathcal{V} \times \mathcal{U}$) *Two continuous intervals $\mathcal{V}$ and $\mathcal{U}$, say $\mathcal{V} = [0, 1]$ and $\mathcal{U} = [0, 1]$, describes a continuous region. Now the RKHS $\mathcal{H}_{\langle u\rangle}$ on the u domain can have the same cubic spline decomposition as that of $\mathcal{H}_{\langle v\rangle}$ in Example 1.1.2, and its tensor product with $\mathcal{H}_{\langle v\rangle}$ defines a type of tensor product cubic splines. One can then build up all the decompositions similar to Example 1.1.2.* □

Example 1.1.4 (Multi-dimensional $\mathcal{U}$) *A multi-dimensional $\mathcal{U}$ corresponds to a multivariate covariate U. Clearly the tensor product structures in Examples 1.1.2 and 1.1.3 can be augmented to accommodate the additional dimensions.* □

1.1.2 Computation

The first term of (1.3) depends on η only through the evaluations $[x_i]\eta = \eta(x_i)$, so the argument of Section 2.3.2 in Gu [2002] applies and the minimizer η_λ of (1.3) has an expression

$$\eta = \sum_{\nu=1}^{m} d_\nu \phi_\nu + \sum_{j=1}^{N} c_j \xi_j = \boldsymbol{\phi}^T \mathbf{d} + \boldsymbol{\xi}^T \mathbf{c}, \tag{1.5}$$

where $\boldsymbol{\phi}$ and $\boldsymbol{\xi}$ are vectors of functions and $\mathbf{d}$ and $\mathbf{c}$ are vectors of coefficients. Straightforward calculation can show that the functional $-\sum_{i=1}^{n}\{Y_i\eta(x_i)-b(\eta(x_i))\}$ is continuous and convex in $\eta \in \mathcal{H}$. So the minimizer η_λ of (1.3) uniquely exists. For a fixed smoothing parameter λ, the minimizer η_λ may be computed via the Newton iteration. Write $\tilde{u}_i = -Y_i + \dot{b}(\tilde{\eta}(x_i)) = -Y_i + \tilde{\mu}(x_i)$ and $\tilde{w}_i = b''(\tilde{\eta}(x_i)) = \tilde{v}(x_i)$. The quadratic approximation of $-Y_i\eta(x_i)+b(\eta(x_i))$ at $\tilde{\eta}(x_i)$ is

$$-\frac{1}{2}\tilde{w}_i\left\{\eta(x_i) - \tilde{\eta}(x_i) + \tilde{u}_i/\tilde{w}_i\right\}^2 + C_i \tag{1.6}$$

where C_i is a number not involving $\eta(x_i)$. The Newton iteration updates $\tilde{\eta}$ by the minimizer of the penalized weighted least squares functional

$$-\frac{1}{n}\sum_{i=1}^{n}\tilde{w}_i(\tilde{Y}_i - \eta(x_i))^2 + \lambda J(\eta) \tag{1.7}$$

where $\tilde{Y}_i = \tilde{\eta}(x_i) - \tilde{u}_i/\tilde{w}_i$. noting that

$$J(\eta) = \Big\langle \sum_{j=1}^{N} c_j\xi_j \,,\, \sum_{k=1}^{N} c_k\xi_k \Big\rangle = \sum_{j=1}^{N}\sum_{k=1}^{N} c_jc_kR_J((X_j^*, U_j^*), (X_k^*, U_k^*)) = \boldsymbol{c}^TQ\boldsymbol{c}, \tag{1.8}$$

The selection of the smoothing parameters can be done through an outer-loop optimization of a cross-validation score. The direct cross-validation was proposed by Cox and Chang [1990]. Xiang and Wahba [1996] derived the more effective and computable GACV score. Gu and Xiang [2001] derived the numerically stable, readily computable GACV score.

1.2 Penalized Likelihood for Lifetime Data

Let T_i be the lifetime of an item, Z_i be the left-truncation time at which the item enters the study, and C_i be the right-censoring time beyond which the item is dropped from the study; T_i and C_i independent of each other. One observes $(Z_i, X_i, \delta_i, U_i)$, $i = 1, \ldots, n$, where $X_i = \min(T_i, C_i)$, $\delta_i = I_{[T_i \le C_i]}$, $Z_i < X_i$, and U_i is a covariate. Assume that $T_i|U_i$ follow a survival function $S(t, u) = \text{Prob}(T > t|U = u)$. We discuss the accelerated life models through location-scale families for the log lifetime. Let $F(z)$ be a cumulative distribution function on $(-\infty, \infty)$ and $f(z)$ be its density. A location-scale family is defined by $P(X \le x|\mu, \sigma) = F((x-\mu)/\sigma)$, where μ is the location parameter and $\sigma > 0$ is the scale parameter. Assume a location-scale family for $\log T$. The survival function and the hazard function are easily seen to be

$$S(t) = 1 - F(z), \quad \lambda(t) = \frac{1}{\sigma t} \frac{f(z)}{1 - F(z)}, \tag{1.9}$$

where $z = (\log t - \mu)/\sigma$. We shall use $\eta = \mu$ from now on. Let σ be a constant and η be a function of a covariate u with $\eta(u_0) = 0$ at a control point u_0. It follows that

$$S(t|u) = 1 - F((\log t\eta(u))/\sigma) = 1 - F(\log(te^{(u)})/\sigma) = S(te^{(u)}|u_0), \tag{1.10}$$

so the covariate is effectively rescaling the time axis. Such models are known as accelerated life models. For example, if we set $F(z) = 1 - e^{-\omega}$ with $f(z) = \omega e^{-\omega}$, where $\omega = e^z$, we

have the extreme value distribution. When $\log T$ follows the extreme value distribution, T follows the Weibull distribution with the survival function and the hazard function

$$S(t) = \exp\{-e^{(\log t-\eta)/\sigma}\} = \exp\{-(t/\beta)^{1/\sigma}\} = \exp\{-(t/\beta)^{\nu}\}, \tag{1.11}$$

$$\lambda(t) = \frac{1}{\sigma t}e^{(\log t-\eta)/\sigma} = \frac{1}{\sigma t}(\frac{t}{e^{\eta}})^{1/\sigma} = \frac{\nu}{t}(\frac{t}{\beta})^{\nu}, \tag{1.12}$$

where $\nu = 1/\sigma$ is called the shape parameter and $\beta = e^{\eta}$ is called the scale parameter. When $\nu = 1$, the Weibull distribution reduces to the exponential distribution.

The minus log likelihood of (Z, X, δ) is seen to be

$$-\{\delta \log \lambda(X; \eta, \sigma) - \int_{Z}^{X} \lambda(t; \eta, \sigma)dt\} = l(\eta, \sigma), \tag{1.13}$$

where λ has a specific form and we want to estimate $\eta(\boldsymbol{u})$ and σ through minimizing

$$-\frac{1}{n}\sum_{i=1}^{n}\{\delta_i \log \lambda(X_i; \eta(\boldsymbol{U_i}), \sigma) - \int_{Z_i}^{X_i} \lambda(t; \eta(\boldsymbol{U_i}), \sigma)dt\} + \frac{\lambda}{2}J(\eta). \tag{1.14}$$

When T follows a Weibull distribution, (1.14) becomes

$$-\frac{1}{n}\sum_{i=1}^{n}\{\delta_i[\nu(\log X_i - \eta(\boldsymbol{U}_i)) + \log \nu - \log X_i] - (X_i^{\nu} - Z_i^{\nu})e^{-\nu\eta(\boldsymbol{U}_i)}\} + \frac{\lambda}{2}J(\eta), \tag{1.15}$$

and we may use the quadratic approximation of $\eta(\boldsymbol{U}_i)$ at $\tilde{\eta}(\boldsymbol{U}_i)$ to obtain a penalized weighted least squares functional similar to (1.7) and iterate on it to calculate the min-

imizer η_λ. ν is also updated in each iteration by maximizing the log likelihood without penalty assuming fixed η after the last iteration.

1.3 The EM algorithm and Louis' formula

1.3.1 The EM algorithm

The EM algorithm was first introduced in the classic 1977 paper by Dempster, Laird, and Rubin Dempster et al. [1977]. They pointed out that the method had been "proposed many times in special circumstances" by other authors, but the 1977 paper generalized the method and developed the theory behind it.

Their paper presents a general approach to iterative computation of maximum-likelihood estimates when the observations can be viewed as incomplete data. Since each iteration of the algorithm consists of an expectation step followed by a maximization step they call it the EM algorithm. The algorithm is particularly useful when the original likelihood function is difficult to optimize but the complete likelihood with the introduction of latent variables is much easier to handle. Based on remarkably simple and general theory, the EM procedure has a wide range of applications, ranging from standard incomplete data problems (e.g. censored and truncated), to iteratively reweighted least squares analysis and empirical Bayes models.

First, consider a general situation. Given a likelihood function $L(\theta; x, y)$, where θ is the parameter vector, x is the observed data and y represents the unobserved latent data or

missing values, the maximum likelihood estimate (MLE) is defined as the maximizer of the marginal likelihood of the observed data $L(\theta; x) = \int_{\mathcal{Y}} L(\theta; x, y)dy$, where $\mathcal{Y}$ is the domain of y. However $L(\theta; x)$ is often intractable. Suppose that $\theta^{(t)}$ denotes the current value of θ after t iterations of the algorithm. The EM algorithm seeks to find the MLE by iteratively applying the following two steps:

Expectation step: Calculate the expected value of the log likelihood function, with respect to the conditional distribution of y given x under the current estimate of the parameters $\theta^{(t)}$:

$$Q(\theta|\theta^{(t)}) = E_{Y|x,\theta^{(t)}} \left[\log L(\theta; x, Y)\right]$$

Maximization step: Find the parameter which maximizes this quantity:

$$\theta^{(t+1)} = \underset{\theta}{\text{argmax}} Q(\theta|\theta^{(t)})$$

Now, consider a special exponential family case. Suppose that $f(x, y|\theta)$ has the regular exponential family form

$$f(x, y|\theta) = b(x, y) \exp\{\theta t(x, y)^T\}/a(\theta) \tag{1.16}$$

where θ denotes a $1 \times r$ vector parameter, $b(x, y)$ denotes a $1 \times r$ vector of complete-data sufficient statistics. The term regular means here that θ is restricted only to an r-dimensional convex set Ω such that the density for all θ in Ω. The parameterization θ is thus unique up

to an arbitrary non-singular $r \times r$ linear transformation, as is the corresponding choice of $b(x, y)$. Such parameters are often called natural parameters, although in familiar examples the conventional parameters are often non-linear functions of θ. The cycle can be described in two steps, as follows:

E-step: Estimate the complete-data sufficient statistics $b(x, y)$ by finding

$$b^{(t)} = E(t(x, y)|x, \theta^{(t)})$$

M-step: Determine $\theta^{(t+1)}$ as the solution of the equations

$$E(b(x, y)|\theta) = b^{(t)}$$

Although an EM iteration does not decrease the observed data likelihood function, there is no guarantee that the sequence converges to a maximum likelihood estimator. For multimodal distributions, this means that an EM algorithm may converge to a local maximum of the observed data likelihood function, depending on starting values. There is a variety of heuristic or metaheuristic approaches for escaping a local maximum such as random restart (starting with several different random initial estimates $\theta^{(t)}$), or applying simulated annealing methods.

EM is particularly useful when the likelihood is an exponential family: the E-step becomes the sum of expectations of sufficient statistics, and the M-step involves maximizing a linear function. In such a case, it is usually possible to derive closed form updates for each step.

An EM algorithm can be easily modified to find the maximum a posteriori (MAP) estimates for Bayesian inference.

There are other methods for finding maximum likelihood estimates, such as gradient descent, conjugate gradient or variations of the Gauss-Newton method. Unlike EM, such methods typically require the evaluation of first and/or second derivatives of the likelihood function.

EM is frequently used for data clustering in machine learning and computer vision. In natural language processing, two prominent instances of the algorithm are the Baum-Welch algorithm (also known as forward-backward) and the inside-outside algorithm for unsupervised induction of probabilistic context-free grammars. The EM algorithm (and its faster variant OS-EM) is also widely used in medical image reconstruction, especially in positron emission tomography and single photon emission computed tomography.

A number of methods have been proposed to accelerate the sometimes slow convergence of the EM algorithm, such as those utilising conjugate gradient and modified Newton Raphson techniques; see, e.g., Jamshidian and Jennrich [1997] and Liu et al. [1998]. Additionally EM can be used along with constrained estimation techniques. Expectation conditional maximization (ECM) replaces each M-step with a sequence of conditional maximization (CM)

steps in which each parameter θ is maximized individually, conditionally on the other parameters remaining fixed Meng and Rubin [1993]. This idea is further extended in generalized expectation maximization (GEM) algorithm, in which one only seeks an increase in the objective function F for both the E step and M step under the alternative description Neal et al. [1999].

EM is a partially non-Bayesian, maximum likelihood method. Its final result gives a probability distribution over the latent variables (in the Bayesian style) together with a point estimate for θ (either a maximum likelihood estimate or a posterior mode). We may want a full Bayesian version of this, giving a probability distribution over θ as well as the latent variables. In fact the Bayesian approach to inference simply treats θ as another latent variable. In this paradigm, the distinction between the E and M steps disappears. We may iterate over each latent variable (now including θ) and optimize them one at a time. There are now k steps per iteration, where k is the number of latent variables.

1.3.2 Louis' formula

The primary conceptual power of the iterative EM algorithm lies in converting a maximization problem involving a complicated likelihood, into a sequence of "pseudo-complete" problems, where at each step the updated parameter estimates can be obtained in a closed form (or at least in a straightforward manner). Unlike Newton-Raphson or Fletcher-Powell techniques, no gradients or curvature matrices need to be derived. Unfortunately this con-

ceptual and analytic simplification does not appear to provide a means of estimating the information matrix associated with the maximum likelihood estimates. There have been, however, solutions published for a few special cases.

Louis' formula Louis [1982] offers a technique for computing the observed information within the EM framework. It requires computation of the complete data gradient and second derivative matrix and can be embedded quite simply in the EM iterations. Of course, an alternative to Louis' formula is use of a derivative-free function maximizing algorithm. In addition, Louis' formula can be used in a way to improve the speed of convergence of the EM algorithm.

We assume that the regularity conditions in Zacks [1971] hold. These guarantee that the MLE solves the gradient equation and that the Fisher information exists. To see how to compute the observed information in the EM, let $S(x, y, \theta) = dl(x, y, \theta)/d\theta$ and $S^*(x, \theta) = dl(x, \theta)/d\theta$ be the gradient vectors of log complete likelihood and log observed likelihood respectively and $B(x, y, \theta) = -d^2l(x, y, \theta)/d\theta^2$ and $B^*(x, \theta) = -d^2l(x, \theta)/d\theta^2$ be the negatives of the associated second derivative matrices. Then by straightforward differentiation:

$$
\begin{aligned}
&S^*(x, \theta) = E_\theta[S(x, Y, \theta)], \\
&S^*(x, \widehat{\theta}) = 0, \\
&I_{obs}(\theta) = E_\theta[B(x, Y, \theta)] - E_\theta[S(x, Y, \theta)S(x, Y, \theta)^T] + S^*(x, \theta)S^{*T}(x, \theta) \qquad (1.17)
\end{aligned}
$$

The first term in (1.17) is the conditional expected full data observed information matrix, while the last two produce the expected information for the conditional distribution of (x, y) given x. That is, using a simplified notation:

$$I_{obs} = I_{full} - I_{full|obs}$$

which is an application of the missing information principle Woodbury [1971] to the observed information. Notice that all of these conditional expectations can be computed in the EM algorithm using only S and B, which are the gradient and curvature for a complete-data problem. Of course, they need be evaluated only on the last iteration of the EM procedure, where S^* is zero.

Example 1.3.1 (Multinomial distribution:) *Here, θ is to be estimated from the multinomial distribution:*

$$\{(0.5 + 0.25\theta), 0.25(1-\theta), 0.25(1-\theta), 0.25\theta\},\ 0 \leq \theta \leq 1.$$

With Y_1, Y_2, Y_3, Y_4 as the frequencies, let

$$Y_1 = X_1 + X_2,\ Y_2 = X_3,\ Y_3 = X_4,\ Y_4 = X_5,$$

where X is multinomial with parameters

$$\{0.5, 0.25\theta, 0.25(1-\theta), 0.25(1-\theta), 0.25\theta\}.$$

Therefore, if X were observed,

$$\widehat{\theta} = \frac{X_2 + X_5}{X_2 + X_3 + X_4 + X_5}.$$

The MLE can be found by solving a quadratic and with data $Y = (125, 18, 20, 34)$, $\widehat{\theta} =$ 0.6268215... . Alternatively, the EM algorithm can be used where

$$X_2^{(t+1)} = \frac{0.25\theta^{(t)}}{0.5 + 0.25\theta^{(t)}},\ Y_1 = \frac{\theta^{(t)}}{2 + \theta^{(t)}} Y_1,\ X_1^{(t+1)} = Y_1 - X_2^{(t+1)},$$

and $X_3 = Y_2,\ X_4 = Y_3,\ X_5 = Y_4$. Here

$$\theta^{(t+1)} = \frac{Y_1\theta^{(t)} + (2 + \theta^{(t)})Y_4}{Y_1\theta^{(t)} + (2 + \theta^{(t)})(Y_2 + Y_3 + Y_4)} = \frac{X_2^{(t+1)} + X_5}{X_2^{(t+1)} + X_3 + X_4 + X_5}.$$

Here

$$S(X, \theta) = \frac{X_2 + X_5}{\theta} + \frac{X_3 + X_4}{1 - \theta},$$

$$B(X, \theta) = \frac{X_2 + X_5}{\theta^2} + \frac{X_3 + X_4}{(1 - \theta)^2}.$$

□

Efron and Hinkley [1978] define I_X as the observed information and show that in most cases it is a more appropriate measure of information than the a priori expectation $E_\theta[B^*(X,\theta)]$. It is certainly easier to compute.

2. ACCELERATED FAILURE TIME MODEL WITH NONPARAMETRIC SPLINE ESTIMATED COMPONENTS FOR CURE RATE DATA

2.1 Introduction

Due to rapid developments and significant progress in medical and health sciences, many diseases can be considered as cured if the patients don't develop a recurrence of the disease for a certain period of time. Consequently, the population consists of the susceptible and non-susceptible subpopulations with the latter termed as "cured". To address this issue, we often need survival models that can incorporate a cure rate to account for the subjects not at risk of death (or relapse). The modeling of both the cure probability and the hazard function of the susceptible subpopulation is of practical interest.

A two-component mixture cure rate model was first proposed by Boag[1949] and later by Berkson and Gage [1952] to analyze the proportion of the cured subjects as well as the failure time distribution of the susceptible subjects. This model assumes that the population consists of two subpopulations. A subject either belongs to the subpopulation considered

as permanently cured or to the subpopulation who are still under risk and will eventually develop an event. Thus the survival function for a subject with covariates $\boldsymbol{x}$ and $\boldsymbol{z}$ is

$$S_{\text{pop}}(t|\boldsymbol{x}, \boldsymbol{z}) = 1 - \pi(\boldsymbol{z}) + \pi(\boldsymbol{z}) S_u(t|\boldsymbol{x}), \tag{2.1}$$

where $\pi(\boldsymbol{z})$ is the probability of being uncured, and $S_{\text{pop}}(t|\boldsymbol{x}, \boldsymbol{z})$ and $S_u(t)$ are the survival functions of the failure time of a subject and the survival function of the failure time of an uncured subject respectively; $\boldsymbol{z}$ and $\boldsymbol{x}$ are the associated covariates. One advantage of this model is that the two components of the mixture cure model can depend on different covariates ($\boldsymbol{x}$ and $\boldsymbol{z}$ may or may not be the same), allowing for separate covariate interpretations for the cure rate function and the survival function of the uncured sub-population and resulting in more flexibility in model selections. In contrast, another competing category of cure rate models, promotion time cure models, have only one covariate function that are associated with both the whole population survival function and cure rate. The mixture model-based approach has been the dominant one in the literature on cure models. Farewell [1982] assumed parametric models on both the cure rate function and the survival function for the subpopulation. Kuk and Chen [1992] extended Farewell [1982] to incorporate semi-parametric Cox proportional hazards model in the survival function. Peng and Dear [2000] applied an EM algorithm based on a marginal likelihood approach and Sy and Taylor [2000] implemented the EM algorithm for the estimation through a full likelihood approach. Lu and Ying [2004] used the mixture formulation to extend a class of semiparametric transformation models proposed by Cheng et al. [1995] to incorporate cure fractions with the proportional

hazards cure model and the proportional odds cure model as special cases. In Othus et al. [2009], the semiparametric transformation model that allows for covariates as well as dependent censoring was proposed. In sum, the most common model for $\pi(\boldsymbol{z})$ is parametric logistic regression and those for $S(t|\boldsymbol{x})$ are parametric regression models (Farewell [1982] Yamaguchi [1992]; Peng, Dear, and Denham [1998]) and semiparametric models such as proportional hazards model (Kuk and Chen [1992]; Peng and Dear [2000]; Sy and Taylor [2000]; Fang, Li, and Sun [2005]), accelerated failure time model (Li and Taylor [2002]; Zhang and Peng [2007]), and semiparametric transformation model (Lu and Ying [2004]; Othus, Li, and Tiwari [2009]).

These mixture cure models have a common limitation in that they all model covariate effects in a parametric form whose validity is generally not justified in practice. The strict parametric assumption can be particularly problematic at the exploratory stage of a study. Recently Wang et al. [2012] developed a family of nonparametric cure rate models based on the smoothing splines ANOVA technique. They assumed the susceptible component has a proportional hazard structure (which means there is no interaction between the time effect and the covariate effect in the log hazard function) to guarantee the identifiability of the mixture model, yielding $S_u(t|\boldsymbol{x}) = S_{u0}(t)^{r(\boldsymbol{x})}$ and the model then becomes

$$S_{\text{pop}}(t|\boldsymbol{x}, \boldsymbol{z}) = \pi(\boldsymbol{z}) S_{u0}(t)^{r(\boldsymbol{x})} + 1 - \pi(\boldsymbol{z}), \tag{2.2}$$

where $S_{u0}(t)$ is the baseline survival function of uncured subjects; $S_{u0}(t)$ and $r(\boldsymbol{x})$ are of purely nonparametric form. If we use $h_{u0}(t)$ to denote the baseline hazard function corresponding to $S_{u0}(t)$ in model (2.2). The hazard function for the non-cured part is $h(t|\boldsymbol{x}) = h_{u0}(t)r(\boldsymbol{x})$, which is quite similar to the classic proportional hazards model (PH model) of Cox [1972], except that $h_{u0}(t)$ and $r(\boldsymbol{x})$ have nonparametric forms in this situation. Despite its popularity, the Cox model structure is often criticized as lack of physical interpretation, especially when comparing to its competitor, accelerated lifetime models, which do not exhibit proportional hazards and describe a situation where the biological or mechanical life history of an event is accelerated or decelerated due to covariate effects. The inventor of the model, Sir D.R. Cox, noted that biological interpretation of the proportional hazards assumption can be quite tricky, and he even commented that "accelerated life models are in many ways more appealing" than the proportional hazards model "because of their quite direct physical interpretation"; see, e.g., Reid [1994], Cox [1997]. This motivates us to find an alternative to the PH model structure. In fact, using PH model is not the only way to specify the effects of $\boldsymbol{x}$ on $S(t)$. Other options include accelerated failure time model (AFT) and accelerated hazard model (AH).

Accelerated hazard model was first proposed by Chen and Wang [2000] without considering a cure fraction of the survival data, and it is useful in the situation where the effect of the covariates is gradually released on the failure time distribution; see, e.g., Zucker and Karr [1990]. The AH model specifies the effect of the covariate by $S(t|\boldsymbol{x}) = S_0(te^{\beta^T x})^{exp(-\beta^T x)}$. Zhang and Peng [2009] put the AH model in the mixture cure model by assuming the AH

model for covariate effects on the failure time distribution of uncured subjects, and they proposed a semi-parametric rank-based method to estimate the parameters in the cure model. Although useful in some cure rate studies, the form of the model makes both the extension to nonparametric estimation and the interpretation complicated. Thus we don't pursue this direction in this thesis.

Accelerated failure time model (AFT model) is a model that provides an alternative to the commonly used proportional hazards models with the advantage of the results being more easily interpretable, and the AFT model postulates a direct relationship between failure time and covariates. This gives us an option to substitute $S_{u0}(t)^{r(\boldsymbol{x})}$ in the PH model with an AFT term $S_{u0}(te^{r(\boldsymbol{x})})$, in which the treatment effect or the covariate $\boldsymbol{x}$ is identified as a time scale change between survival functions. The AFT model as a semi-parametric family has received considerable attention in statistical literature, and our interest is to extend the semi-parametric AFT models with parametric form of covariate effects to nonparametric modeling of covariate effects under the AFT setting.

2.2 The Model

Let $(t_i, \delta_i, \boldsymbol{z_i}, \boldsymbol{x}_i)$ be the observed data for the ith subject, $i = 1, \cdots, n$. Here t_i is the observed lifetime time for the ith subject, δ_i is an indicator with $\delta_i = 1$ for observed failures and $\delta_i = 0$ for censored subjects, and $\boldsymbol{z}_i$, $\boldsymbol{x}_i$ are the covariates. Note that all the cured subjects are censored and have $\delta_i = 0$, but that some censored subjects may also experience

failures after the study. Assuming independent and non-informative censoring, the observed likelihood function can be written as

$$l_{obs}(\pi(\cdot), S_u(\cdot)) = \prod_{i=1}^{n} [\pi(\boldsymbol{z}_i) f_u(t_i|\boldsymbol{x}_i)]^{\delta_i} [\pi(\boldsymbol{z}_i) S_u(t_i|\boldsymbol{x}_i) + 1 - \pi(\boldsymbol{z}_i)]^{1-\delta_i}, \tag{2.3}$$

where $f_u(t, \boldsymbol{x})$ and $S_u(t, \boldsymbol{x})$ are respectively the probability density function and the survival function of failure time t given the covariate $\boldsymbol{x}$. Assuming a location-scale family distribution for $\log T$, the proposed accelerated lifetime model considers a nonparametric model for the location parameter η such that it is a function $\eta(\boldsymbol{x})$ of the covariate $\boldsymbol{x}$. Then $S_u(t, \boldsymbol{x}) = S_{u\sigma}(te^{-\eta(\boldsymbol{x})})$, where $S_{u\sigma}(t) = S_{u0}(t/\sigma)$ is the so-called baseline survival function known up to the constant scale parameter σ and S_{u0} is the known base survival function for the location-scale family. Note that the covariate effect is to change the time scale by a factor $e^{-\eta(\boldsymbol{x})}$, either accelerating or decelerating the time, which earns its name. The corresponding hazard rate function is $h_u(t, \boldsymbol{x}) = e^{-\eta(\boldsymbol{x})} h_{u\sigma}(te^{-\eta(\boldsymbol{x})})$, where $h_{u\sigma}$ is the hazard rate function corresponding to S_σ. To emphasize its dependence on η and σ, we will use $h(t, \boldsymbol{x}; \eta, \sigma)$ to denote $h(t, \boldsymbol{x})$ from now on.

Among many choices in the AFT model term that satisfy the structure $S_0(t|\boldsymbol{x}) = S_0(te^{r(\boldsymbol{x})})$, such as log-logistic, Weibull, log-normal, gamma and inverse Gaussian distributions, we adopt the Weibull distribution as our first choice since Weibull distribution can be parameterised as either a proportional hazards model or an AFT model, and therefore the results of fitting a Weibull model can be interpreted in either framework. For the Weibull distribu-

tion, we have $S_{u0}(x) = \exp(-x^\nu)$ and $r(x) = -\eta(x)$. The survival function for the non-cured subpopulation then becomes $S_{u0}(te^{r(\boldsymbol{x})}) = \exp\{-(\frac{t}{e^{\eta(\boldsymbol{x})}})^\nu\}$, where $\eta(\boldsymbol{x})$ and ν are the scale parameter and the shape parameter of the Weibull distribution respectively.

Assuming Weibull distribution for AFT, the mixture cure model is then

$$S_{\text{pop}}(t|\boldsymbol{x},\boldsymbol{z}) = \pi(\boldsymbol{z})\exp\{-(\frac{t}{e^{\eta(\boldsymbol{x})}})^\nu\} + 1 - \pi(\boldsymbol{z}), \tag{2.4}$$

where ν is an unknown shape parameter and $\eta(\boldsymbol{x})$ is an unknown smooth function.

Before developing the estimation procedure,, an important issue is to ensure that such a model is identifiable. We need to make sure that for any $S_{\text{pop}}(t|\boldsymbol{x},\boldsymbol{z}) = \tilde{S}_{\text{pop}}(t|\boldsymbol{x},\boldsymbol{z})$, $(\pi(\boldsymbol{z}),\eta(\boldsymbol{x}))$ and $(\tilde{\pi}(\boldsymbol{z}),\tilde{\eta}(\boldsymbol{x}))$ satisfying the equation are exactly the same functions. It can be proved that the model is identifiable when ν is fixed. The proof of the identifiability of the model is at the end of this chapter.

2.3 Penalized EM Method

Suppose that the data is of the form $(t_i,\delta_i,\boldsymbol{x}_i,\boldsymbol{z}_i)$, the observed-data likelihood function of the parameters $(\pi(\boldsymbol{z}),\eta(\boldsymbol{x}))$ for the mixture model (the incomplete log likelihood) is then given by

$$L = \prod_{i=1}^{n}[\pi(\boldsymbol{z}_i)f(t_i|\boldsymbol{x}_i)]^{\delta_i}[\pi(\boldsymbol{z}_i)S(t_i|\boldsymbol{x}_i) + 1 - \pi(\boldsymbol{z}_i)]^{1-\delta_i}, \tag{2.5}$$

where $f(t_i|\boldsymbol{x}_i)$ is the density function of the survival part. It is difficult to directly optimize the likelihood function and we employ EM algorithm by introducing a latent variable Z_i. Let Z_i be the indicator of the cure status of the ith patient such that $Z_i = 1$ if the ith patient is not cured and $Z_i = 0$ if cured. Obviously, if $\delta_i = 1$, then $Z_i = 1$. Z_i is a latent variable and cannot be observed directly, but it plays an important role in the EM algorithm as bridge between the two parts, the expectation part and the maximization part. Given Z_i, t_i and δ_i, the complete log likelihood function under the model assumption now changes to

$$L_c = \log \prod_{i=1}^{n} \left\{ [\pi(\boldsymbol{z}_i) f(t_i|\boldsymbol{x}_i)]^{Z_i} \right\}^{\delta_i} \cdot \left\{ [\pi(\boldsymbol{z}_i) S(t_i|\boldsymbol{x}_i)]^{Z_i} [1 - \pi(\boldsymbol{z}_i)]^{1-Z_i} \right\}^{1-\delta_i}. \tag{2.6}$$

The Expectation-step is to calculate the estimated expectation of log likelihood function L_c, given the estimated components $\tilde{S}$ and $\tilde{\pi}$. From the model assumption we can derive the relationship between $E(Z_i)$ and $(\pi(\boldsymbol{z}_i), S(t_i|x_i))$, which is

$$E(Z_i) = \delta_i + (1 - \delta_i) \frac{\pi(\boldsymbol{z}_i) S(t_i|x_i)}{1 - \pi(\boldsymbol{z}_i) + \pi(\boldsymbol{z}_i) S(t_i|x_i)}. \tag{2.7}$$

The expectation $E(L_c)$ is the sum of the following two functions,

$$\begin{aligned} E(L_1) &= \sum_{i=1}^{n} \Big(E(Z_i) \log \pi(\boldsymbol{z}_i) + \big(1 - E(Z_i)\big) \log \big(1 - \pi(\boldsymbol{z}_i)\big) \Big) \\ E(L_2) &= \sum_{i=1}^{n} \Big(E(Z_i) \log S(t_i|\boldsymbol{x}_i) + \delta_i E(Z_i) \log \lambda(t_i|\boldsymbol{x}_i) \Big), \end{aligned} \tag{2.8}$$

where $\lambda(\cdot)$ is the hazard function of the failure time distribution of the uncured part. The constraint that a non-cure rate function must satisfy is that $\pi(\boldsymbol{z}) \in [0,1]$ on $\mathcal{Z}$. One can make a one-to-one logit transform $\zeta(z) = \log \frac{\pi(z)}{1-\pi(z)}$ and estimate ζ instead, which is free of the positivity and upper bound constraints. And particularly for the Weibull distribution in our application, the survival function is $S(t|x) = e^{-(\frac{t}{e^{\eta(x)}})^{\nu}}$ and the hazard function is $\lambda(t|x) = \frac{\nu}{e^{\eta(x)}}(\frac{t}{e^{\eta(x)}})^{\nu-1}$. So the expectation of log likelihood functions become

$$\begin{aligned} E[L_1(\zeta; \boldsymbol{Z})] &= \sum_{i=1}^{n} \left\{ E(Z_i)\zeta(\boldsymbol{z}_i) - \log[1 + e^{\zeta(\boldsymbol{z}_i)}] \right\} \\ E[L_2(\eta, \nu; \boldsymbol{Z})] &= \sum_{i=1}^{n} \left\{ \delta_i[\nu(\log t_i - \eta(\boldsymbol{x}_i)) + \log \nu - \log t_i] - E(Z_i) t_i^{\nu} e^{-\nu\eta(\boldsymbol{x}_i)} \right\}. \end{aligned} \tag{2.9}$$

Thus if we know the estimated $\tilde{S}(t_i|x_i)$ and $\tilde{\pi}(\boldsymbol{z}_i)$ from the last maximization step, we can estimate the expectation of Z_i according to the Bayesian rule, and then we can update the expectation of the log likelihood functions L_1 and L_2 by using the newly updated $\tilde{E}(Z_i)$.

In the maximization step, our task is to maximize the expectation of the complete log likelihood L_c, which is equivalent to maximize $E(L_1)$ and $E(L_2)$ separately since $E(L_1)$ doesn't involve any $S(t_i|x_i)$ and $\pi(\boldsymbol{z}_i)$ doesn't appear in L_2. We use nonparametric estimation method and allow $S(t|x)$ and $\pi(\boldsymbol{z})$ to vary in high-dimensional function spaces to avoid possible model misspecification in a parametric analysis, so it is necessary to add penalty terms in the maximization step to penalize the roughness of $S(t|x)$ and $\pi(\boldsymbol{z})$ and we can

tune the penalty parameters to control the trade-off between the two conflicting goals, the fitness and the smoothness. The M-step then minimizes

$$-\frac{1}{n}E[L_1(\zeta)] + \frac{\beta}{2}J(\zeta), \tag{2.10}$$

with respect to ζ, and

$$-\frac{1}{n}E[L_2(\eta,\nu)] + \frac{\lambda}{2}J(\eta), \tag{2.11}$$

with respect to η and ν, where $e^{\eta(x)}$ is the scale parameter and ν is the shape parameter in the Weibull distribution. The penalty in (2.10) is seen to be $\frac{\beta}{2}J(\zeta) = \frac{\beta}{2}\sum_{\gamma=1}^{p}\theta_\gamma^{-1}(\zeta,\zeta)_\gamma$, with β and θ_γ as smoothing parameters. $((\cdot,\cdot)_\gamma$ are inner products for subspaces $\mathcal{H}_\gamma$ and a multiple-term reproducing kernel Hilbert space can be written as $\mathcal{H} = \oplus_\gamma \mathcal{H}_\gamma$.)

2.4 Inference

We build the confidence interval using the observed information matrix by employing the Louis missing information principle ($I_{obs} = I_{full} - I_{full|obs}$). To extract the observed information matrix in terms of complete log likelihood, Louis [1982] gives

$$I_{obs}(\theta) = E_\theta[B(x,Y,\theta)] - E_\theta[S_c(x,Y,\theta)S_c(x,Y,\theta)^T] + S(x,Y,\theta)S(x,Y,\theta)^T, \tag{2.12}$$

where S_c is the gradient vector of the complete log likelihood and B is the negative of the second derivative matrix; θ is the unknown parameter vector; Y is the missing data.

In our case, the penalized complete log likelihood is:

$$L_c(\boldsymbol{Z};(\zeta,\eta,\nu)) = L_1(\zeta;\boldsymbol{Z}) - \frac{n\beta}{2}J(\zeta) + L_2(\eta,\nu;\boldsymbol{Z}) - \frac{n\lambda}{2}J(\eta) \tag{2.13}$$

where $L_1(\zeta;\boldsymbol{Z})$ and $L_2(\eta,\nu;\boldsymbol{Z})$ are defined in (2.9). If we write $\zeta(\boldsymbol{z}) = \boldsymbol{\phi}_\zeta^T(\boldsymbol{z})\mathbf{d}_\zeta + \boldsymbol{\xi}_\zeta^T(\boldsymbol{z})\mathbf{c}_\zeta := \psi_\zeta(\boldsymbol{z})^T\mathbf{b}_\zeta$ and $\eta(\boldsymbol{x}) = \boldsymbol{\phi}_\eta^T(\boldsymbol{x})\mathbf{d}_\eta + \boldsymbol{\xi}_\eta^T(\boldsymbol{x})\mathbf{c}_\eta := \psi_\eta(\boldsymbol{x})^T\mathbf{b}_\eta$, then the first derivative and the second derivative of the penalized complete log likelihood are correspondingly,

$$\begin{aligned} S_c(\boldsymbol{Z},\boldsymbol{b}_\zeta,\boldsymbol{b}_\eta,\nu) &= \begin{pmatrix} \frac{\partial L_c}{\partial \boldsymbol{b}_\zeta} \\ \frac{\partial L_c}{\partial \boldsymbol{b}_\eta} \\ \frac{\partial L_c}{\partial \nu} \end{pmatrix} \\ &= \begin{pmatrix} \sum_{i=1}^n \left[Z_i \boldsymbol{\psi}_\zeta(z_i) - \left(1+\exp\left\{-\boldsymbol{\psi}_\zeta(z_i)^T\boldsymbol{b}_\zeta\right\}\right)^{-1}\boldsymbol{\psi}_\zeta(z_i)\right] - n\beta Q_\zeta^*\boldsymbol{b}_\zeta \\ \sum_{i=1}^n \left[-\delta_i\nu\boldsymbol{\psi}_\eta(x_i) + Z_i\nu {t_i}^\nu \exp\left\{-\nu\boldsymbol{\psi}_\eta(x_i)^T\boldsymbol{b}_\eta\right\}\boldsymbol{\psi}_\eta(x_i)\right] - n\lambda Q_\eta^*\boldsymbol{b}_\eta \\ \sum_{i=1}^n \left\{\delta_i\left[\log t_i - \eta(x_i) + \frac{1}{\nu}\right] - Z_i t_i^\nu e^{-\nu\eta(x_i)}\left[\log(t_i) - \eta(x_i)\right]\right\} \end{pmatrix} \end{aligned} \tag{2.14}$$

$$B(\boldsymbol{Z}, \boldsymbol{b}_\zeta, \boldsymbol{b}_\eta, \nu) = \begin{pmatrix} -\frac{\partial^2 L_c}{\partial \boldsymbol{b}_\zeta \partial \boldsymbol{b}_\zeta^T} & 0 & 0 \\ 0 & -\frac{\partial^2 L_c}{\partial \boldsymbol{b}_\eta \partial \boldsymbol{b}_\eta^T} & 0 \\ 0 & 0 & -\frac{\partial^2 L_c}{\partial \nu^2} \end{pmatrix} \tag{2.15}$$

where

$$\begin{aligned} -\frac{\partial^2 L_c}{\partial \boldsymbol{b}_\zeta \partial \boldsymbol{b}_\zeta^T} &= \sum_{i=1}^{n} \left(1 + exp\left\{\boldsymbol{\psi}_\zeta(z_i)^T \boldsymbol{b}_\zeta\right\}\right)^{-2} exp\left\{\boldsymbol{\psi}_\zeta(z_i)^T \boldsymbol{b}_\zeta\right\} \boldsymbol{\psi}_\zeta(z_i) \boldsymbol{\psi}_\zeta(z_i)^T + n\beta Q_\zeta^* \\ -\frac{\partial^2 L_c}{\partial \boldsymbol{b}_\eta \partial \boldsymbol{b}_\eta^T} &= \sum_{i=1}^{n} Z_i t_i^\nu \nu^2 \exp\left\{-\nu \boldsymbol{\psi}_\eta(x_i)^T \boldsymbol{b}_\eta\right\} \boldsymbol{\psi}_\eta(x_i) \boldsymbol{\psi}_\eta(x_i)^T + n\lambda Q_\eta^* \\ -\frac{\partial^2 L_c}{\partial \nu^2} &= \sum_{i=1}^{n} \left\{\delta_i \nu^{-2} + Z_i t_i^\nu e^{-\nu\eta(x_i)} \left[\log(t_i) - \eta(x_i)\right]^2\right\} \end{aligned}$$

where Q_ζ^* and Q_η^* are partitioned matrices of appropriate dimensions with the bottom right positions filled by Q defined in (1.8) and the rest by 0.

For $S_c S_c^T$ part, since $E[Z_i y_j] = E[Z_i]E[y_j]$ for $i \neq j$ and $E[Z_i^2] = E[Z_i]$, it follows:

$$\begin{aligned} & E\left[S_c(\boldsymbol{y}, \boldsymbol{b}_\zeta, \boldsymbol{b}_\eta, \nu) S_c^T(\boldsymbol{y}, \boldsymbol{b}_\zeta, \boldsymbol{b}_\eta, \nu)\right] \\ & = E\left[S_c(\boldsymbol{y}, \boldsymbol{b}_\zeta, \boldsymbol{b}_\eta, \nu)\right] E\left[S_c(\boldsymbol{y}, \boldsymbol{b}_\zeta, \boldsymbol{b}_\eta, \nu)\right]^T + \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{12}^T & A_{22} & A_{23} \\ A_{13}^T & A_{23}^T & A_{33} \end{pmatrix}, \end{aligned} \tag{2.16}$$

where

$$A_{11} = \sum_{i=1}^{n} \left(E[Z_i] - E[Z_i]^2\right) \boldsymbol{\psi}_\zeta(z_i)\boldsymbol{\psi}_\zeta(z_i)^T$$

$$A_{12} = \sum_{i=1}^{n} \left(E[Z_i] - E[Z_i]^2\right) \left(\nu {t_i}^\nu \exp\left\{-\nu \boldsymbol{\psi}_\eta(x_i)^T \boldsymbol{b}_\eta\right\}\right) \boldsymbol{\psi}_\zeta(z_i)\boldsymbol{\psi}_\eta(x_i)^T$$

$$A_{13} = \sum_{i=1}^{n} \left(E[Z_i] - E[Z_i]^2\right) \left(-t_i^\nu e^{-\nu\eta(x_i)} \left[\log(t_i) - \eta(x_i)\right]\right) \boldsymbol{\psi}_\zeta(z_i)$$

$$A_{22} = \sum_{i=1}^{n} \left(E[Z_i] - E[Z_i]^2\right) \left(\nu {t_i}^\nu \exp\left\{-\nu \boldsymbol{\psi}_\eta(x_i)^T \boldsymbol{b}_\eta\right\}\right)^2 \boldsymbol{\psi}_\eta(x_i)\boldsymbol{\psi}_\eta(x_i)^T$$

$$A_{23} = \sum_{i=1}^{n} \left(E[Z_i] - E[Z_i]^2\right) \left(-\nu t_i^{2\nu} e^{-2\nu\eta(x_i)} \left[\log(t_i) - \eta(x_i)\right]\right) \boldsymbol{\psi}_\eta(x_i)$$

$$A_{33} = \sum_{i=1}^{n} \left(E[Z_i] - E[Z_i]^2\right) \left(-t_i^\nu e^{-\nu\eta(x_i)} \left[\log(t_i) - \eta(x_i)\right]\right)^2$$

To obtain A_{ij}s we use the updated latent variable values in the last iteration to replace $E[Z_i]$. Then the observed information matrix is straight forward to calculate. To obtain $100(1-\alpha)\%$ point-wise confidence intervals for $\zeta(z_0)$, $\eta(x_0)$ and ν from the observed information matrix, one can compute

$$\begin{pmatrix} \psi_\zeta(\boldsymbol{z}_0)^T \widetilde{\boldsymbol{b}}_\zeta \\ \psi_\eta(\boldsymbol{x}_0)^T \widetilde{\boldsymbol{b}}_\eta \\ \widetilde{\nu} \end{pmatrix} \pm z_{\alpha/2} \left(Diag \left\{ \begin{pmatrix} \psi_\zeta(\boldsymbol{z}_0)^T & 0 & 0 \\ 0 & \psi_\eta(\boldsymbol{x}_0)^T & 0 \\ 0 & 0 & 1 \end{pmatrix} I_{obs}^{-1} \begin{pmatrix} \psi_\zeta(\boldsymbol{z}_0) & 0 & 0 \\ 0 & \psi_\eta(\boldsymbol{x}_0) & 0 \\ 0 & 0 & 1 \end{pmatrix} \right\} \right)^{1/2}$$

corresponding to the three confidence intervals, where $\widetilde{\zeta}(z_0) = \psi_\zeta(\boldsymbol{z}_0)^T\widetilde{\boldsymbol{b}}_\zeta$, $\widetilde{\eta}(x_0) = \psi_\eta(\boldsymbol{x}_0)^T\widetilde{\boldsymbol{b}}_\eta$ and $\widetilde{\nu}$ are the estimates obtained from the PEM algorithm. Care must be taken when I_{obs} is singular. In practice, one may simply perform the Cholesky decomposition of I_{obs} with pivoting, replace the trailing O (if present) by δI with an appropriate value of δ , then proceed as if I_{obs} were of full column rank. The same technique was used in, e.g., Kim and Gu [2004] for handling the singularity of a Hessian matrix.

2.5 Simulations

We conduct simulation studies to evaluate the empirical performance of our proposed methodology for the estimation and the inference. For penalized likelihood regression, the confidence intervals derived from the Bayes models demonstrate a certain frequentist across-the-function coverage property; see, e.g., Wahba [1983], Nychka [1988] and Gu [1992]. Here, we will assess the coverage properties of the intervals derived in Section 2.5.

Our model can be written as

$$S_{\text{pop}}(t|\boldsymbol{x}, \boldsymbol{z}) = \pi(\boldsymbol{z}) \exp\{-(\frac{t}{e^{\eta(\boldsymbol{x})}})^\nu\} + 1 - \pi(\boldsymbol{z}),$$

where $\pi(z)$ is the cure rate function, $e^{\eta(x)}$ is the scale parameter in the Weibull distribution, and ν is the shape parameter.

We consider three test cure rate functions for π and three different test functions for η. Thus, we have $3 * 3 = 9$ different test function settings. We also repeat every setting with sample size $n = 400$ and $n = 800$. So in total we have $3 * 3 * 2 = 18$ simulations.

For the cure rate component, the three test functions all have one continuous covariate z and the same unimodal shape. The three test cure rate functions are:

$$\pi_i(z) = c_i + 0.7 \sin(2(z + 0.6)), \ i = 1, 2, 3$$

where $c_1 = 0.1722$, $c_2 = -0.0278$, $c_3 = -0.2278$ are constants and they differ by a 0.2 increment. The different choices of c_i control the overall probability of cure rate and are chosen to be most representative. The overall cure probability of all subjects for the three functions are 20%, 40% and 60% respectively.

For the survival component, the three test functions all have one continuous covariate x. The three test functions for the scale parameter in the Weibull distribution are:

$$e^{\eta_1(x)} = .3 \times [10^6 x^{11}(1 - x)^6 + 10^4 x^3 (1 - x)^{10}] + 1$$

$$e^{\eta_2(x)} = \frac{2.5}{(1 + 0.5 \sin(2\pi x))^{2.5}}$$

$$e^{\eta_3(x)} = 3.5(20(x + .5)^2 + .01)^{1.4}/100$$

For the shape parameter, we choose $\nu = 2$. Note that all the hazard functions have the proportional hazard structure with the proportional part $\exp(-\eta(x)\nu)$ depending on the covariate x.

For settings of sample size $n = 400$, the covariates z were generated on a grid of 20 equally spaced values over the range $[-0.4, 0.4]$; the covariates x were generated on a grid of 400 equally spaced values over the range $[0, 1]$. Then 20 samples are generated at each covariate point of z: first, 20 samples are randomly classified as either cured or not cured based on the test cure probability functions; then, failure times are randomly generated for the non-cured samples based on the Weibull distribution with test $e^{\eta(x)}$; finally, censoring times were generated for the non-cured samples and the censoring status indicators were recorded. Note that all the cured samples were recorded as being censored. For sample size $n = 800$, the difference are that they are generated on a grid of 32 equally spaced values for z over the range $[-0.4, 0.4]$ and on a grid of 800 equally spaced values for x over the range $[-1, 1]$, then 25 samples are generated at each covariate point of z. The censoring times were generated from Weibull distributions with the parameters chosen in a way so that the observed censoring rate corresponding to the three settings of π_i are 25%, 45% and 65% respectively.

One hundred replicates were generated for each setting. The point-wise 95% confidence intervals were calculated for logit cure rate $\zeta(z)$ on a z grid of size 100 equally spaced on $[-0.4, 0.4]$, for $\eta(x)$ with x fixed at a randomly picked evaluation point x.

Figure 2.1 shows simulation results of sample sizes $n = 400$ with the same settings as Figure 2.2. By comparing the results from sample size $n = 800$ and $n = 400$, we see that the width of confidence interval decreases and the mean function estimate gets more accurate as sample size increases. The empirical point-wise coverage do not seem to be affected much by sample sizes. The point-wise coverage is generally close to the nominal level 0.95 with a bit under-coverage in certain areas. In general, higher curvature implies a rougher curve which is more difficult to be recovered fully by nonparametric smoothing methods. Second, low coverage also seemed to happen at both ends of the data range where data are sparse and information for nonparametric method is dwindling.

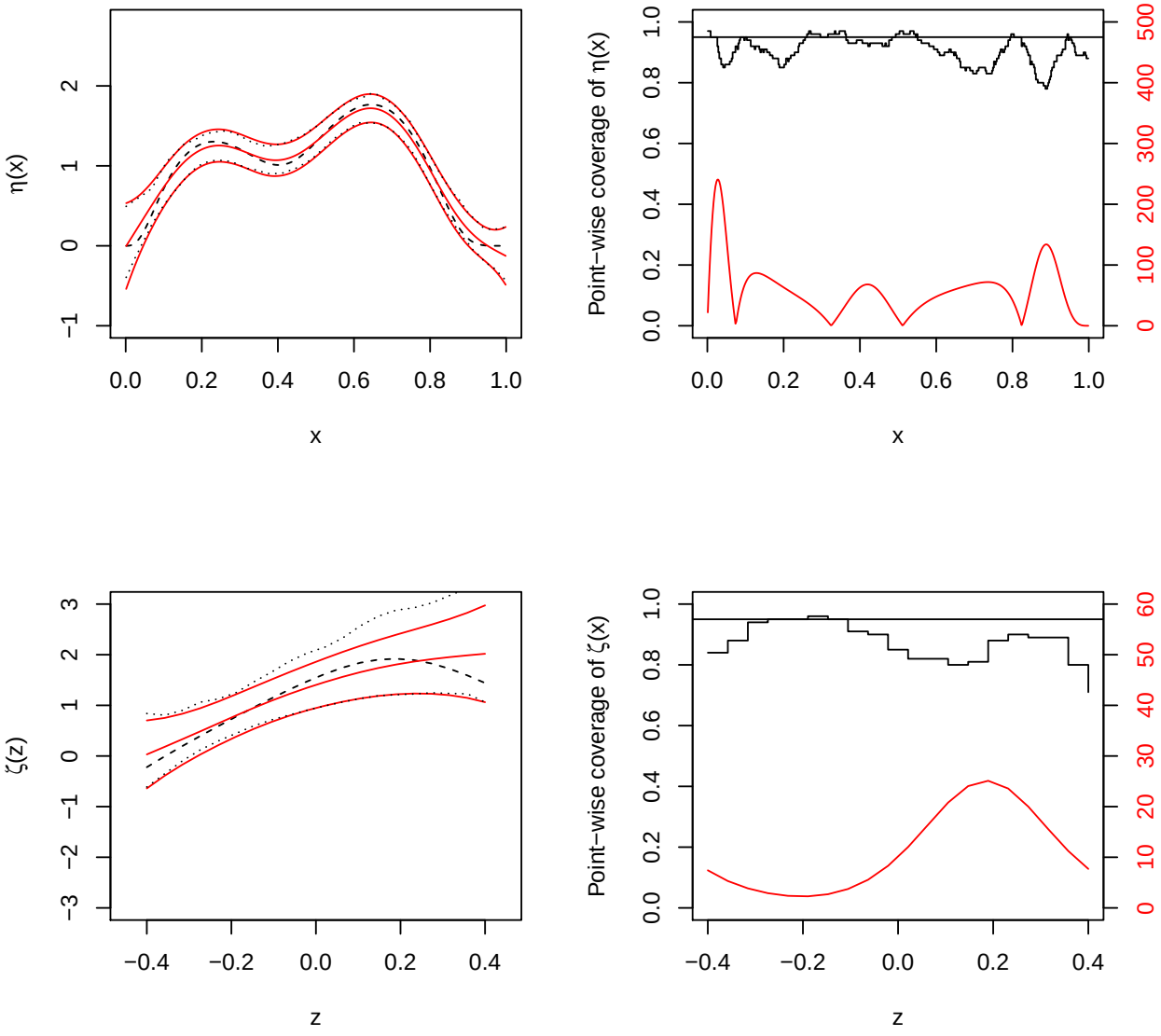

Fig. 2.1. Simulation Results for Test Functions $\eta_1(\boldsymbol{x})$, $\zeta_1(\boldsymbol{z})$ and $n = 400$. Right column: Point-wise coverage (top black lines). Superimposed are nominal coverage (black straight lines) and scaled $|\eta''(x)|$ or $|\zeta''(z)|$. Left column: True functions (dashed) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (solid) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates. The average of the estimated ν is 2.04 (the true $\nu = 2$) and the average of the 95% CIs is $[1.82, 2.27]$. The empirical 2.5 and 97.5 percentiles of the estimated ν is 1.86 and 2.27. The coverage of ν is 0.95.

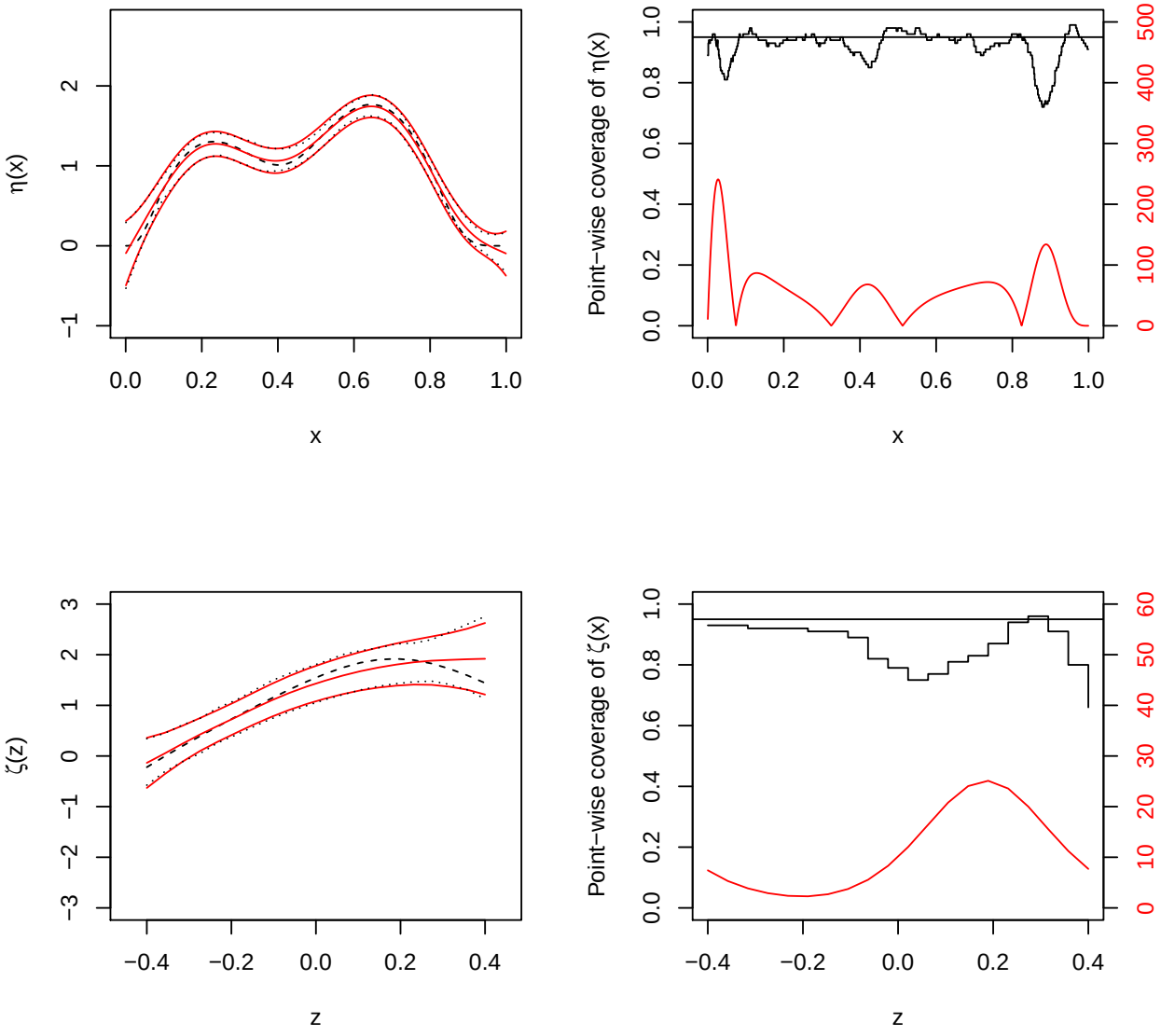

Fig. 2.2. Simulation Results for Test Functions $\eta_1(\boldsymbol{x})$, $\zeta_1(\boldsymbol{z})$ and $n = 800$. Right column: Point-wise coverage (top black lines). Superimposed are nominal coverage (black straight lines) and scaled $|\eta''(x)|$ or $|\zeta''(z)|$. Left column: True functions (dashed) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (solid) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates. The average of the estimated ν is 2.01 (the true $\nu = 2$) and the average of the 95% CIs is $[1.85, 2.16]$. The empirical 2.5 and 97.5 percentiles of the estimated ν is 1.88 and 2.18. The coverage of ν is 0.95.

2.6 Application to Melanoma Data

As an illustration, we applied the two component cure model to a melanoma clinical trial. The proposed method is used to fit the model to a SEER melanoma cancer data set. The Surveillance, Epidemiology and End Results (SEER) Program of the National Cancer Institute is an authoritative source of information on cancer incidence and survival in the United States (http://www.seer.cancer.gov). SEER currently collects and publishes cancer incidence and survival data from population-based cancer registries covering approximately 26% of the US population. SEER coverage includes 23% of African Americans, 40% of Hispanics, 42% of American Indians and Alaska Natives, 53% of Asians and 70% of Hawaiian/Pacific Islanders. The SEER Programme began collecting data on cancer cases in 1973 in the states of Connecticut, Iowa, New Mexico, Utah and Hawaii, and the metropolitan areas of Detroit and San Francisco-Oakland. In 1974 - 1975, the metropolitan area of Atlanta and the 13-county Seattle-Puget Sound area were added. These original nine regions are referred to as the SEER 9 registries, covering 10% of the US population.

Melanoma causes the majority (75%) of deaths related to skin cancer. Worldwide, doctors diagnose about 160,000 new cases of melanoma yearly. According to a WHO report, about 48,000 melanoma-related deaths occur worldwide per year. The treatment includes surgical removal of the tumor. If melanoma is found early, while it is still small and thin, and if it is completely removed, then the chance of cure is high. In fact, many cancer patients live long enough to die ultimately from other causes. This implies that for melanoma cancer a cure

rate model may be helpful to analyze both the cured proportion as well as the hazard rate for the non-cured subpopulation.

The SEER data we consider consist of 637 melanoma cancer cases that satisfy some criteria. These patients didn't have previous other cancer diagnosis before melanoma; they all received routine treatment including surgery and radiotherapy and further they are all white people. Our "failure time" of interest was time from diagnosis of melanoma cancer to death due to melanoma cancer. We consider three covariates, *age*, *gender* and *tumor size*. In women, the most common site is the legs and melanomas in men are most common on the back. This indicates that melanoma may have different mechanism for women and men, so we use the gender as one of the covariates. Relative literatures also show that melanoma is much more dangerous if it is not found in the early stages, so we include the tumor size as another covariate. A question of interest was whether survival or cure fractions differed in this data set by age, gender and tumor size.

A scatter plot of the data is provided in Figure 2.3. The black points represent observed deaths, and the red points represent censored observations. Clearly, there is a large portion of censored observations, especially after 50 months. This may suggests the existence of a subpopulation of cured subjects in the study. Thus a cure rate data analysis is appropriate here.

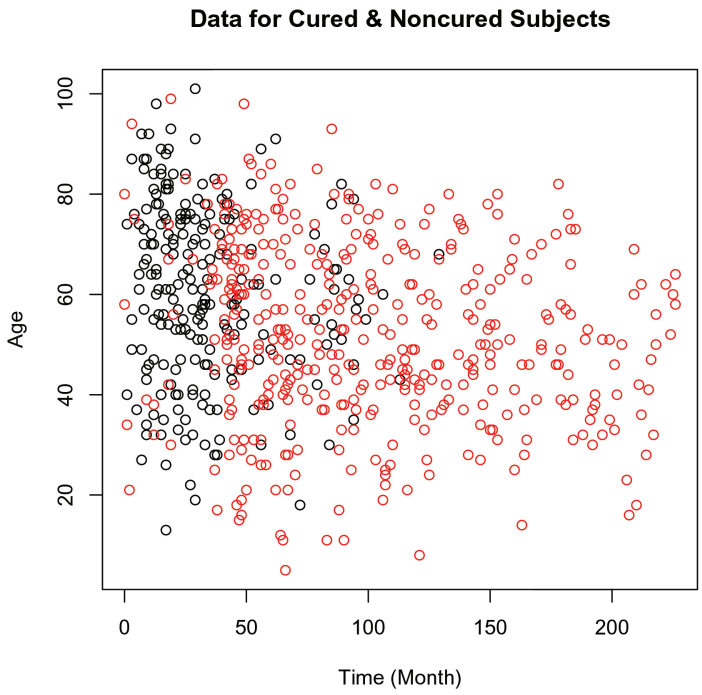

Fig. 2.3. Plot of data. Black circles are observed failures, red circles are observed censoring.

The covariates considered in our example are age at diagnosis with a range of 5 to 101 years, gender (M or F) and tumor size (Big or Small). Both $\boldsymbol{x}$ and $\boldsymbol{z}$ are thus (age, gender, size). We allow interaction between (age, gender, size) for both $\zeta(\boldsymbol{z})$ and $\eta(\boldsymbol{x})$ such that

$$\begin{aligned}\zeta(age, gender, size) = \zeta_0 &+ \zeta_a(age) + \zeta_g(gender) + \zeta_s(size) \\ &+ \zeta_{ag}(age, gender) + \zeta_{as}(age, size) + \zeta_{gs}(gender, size) \\ &+ \zeta_{ags}(age, gender, size)\end{aligned}$$

$$\eta(age, gender, size) = \eta_0 + \eta_a(age) + \eta_g(gender) + \eta_s(size) + \eta_{ag}(age, gender) + \eta_{as}(age, size) + \eta_{gs}(gender, size) + \eta_{ags}(age, gender, size).$$

For the non-cure rate $\pi(\boldsymbol{z})$, Figure 2.4 shows that the CIs for female group do not cover any constant line and the CIs for male group can barely do so. This suggests a likely association of age with the non-cure rate. For male patients, the non-cure rate increases up to age 60 and then levels off; but for female patients, the non-cure rate shows a strong and consistent increasing trend against age. For female group, the cure rates for both sizes of tumors are comparable but the increase of the non-cure rate against age for small size tumors seems to be steeper than the one for large size tumors. An interesting difference is observed between the male and female groups where the patients with big size tumors seem to have a greater non-cure rate than the patients with small size tumors *only* for the male group.

The estimated shape parameter ν for the Weibull distribution for the non-cured subpopulation is 1.45; its estimated standard deviation is 0.0817; the confidence interval for ν is $[1.29, 1.61]$. The estimated log of the scale parameter, $\eta(\boldsymbol{x})$ is illustrated in Figure 2.5.

Based on the estimation and the inference of $\eta(\boldsymbol{x})$ and ν, we can derive the log hazard functions. For the survival component, Figure A.17 and A.18 respectively plot log hazard at $time = 10\ months$ against age and log hazard against time with age fixed at 53 years for the four patient groups. $Time = 10\ months$ is the median of all the failure times and

$age = 53\ years$ is the median of all the ages. Of particular interest is the clear nonlinear trend of log hazard plots against age in Figure A.18. This may suggest a nonlinear form of age effect in the model for log hazard. Hence, a standard proportional hazards model with linear age effect in the log relative risk may not be sufficient to describe the true trend of hazard versus age. The log hazards at $age = 53\ years$ in Figure A.17 all show an increasing trend that is very close to be linear. Also notice that the four plots in Figure A.17 or A.18 are all quite similar to each other. This indicates possibly negligible gender and size effects at the chosen cross-sections of the log hazard surface.

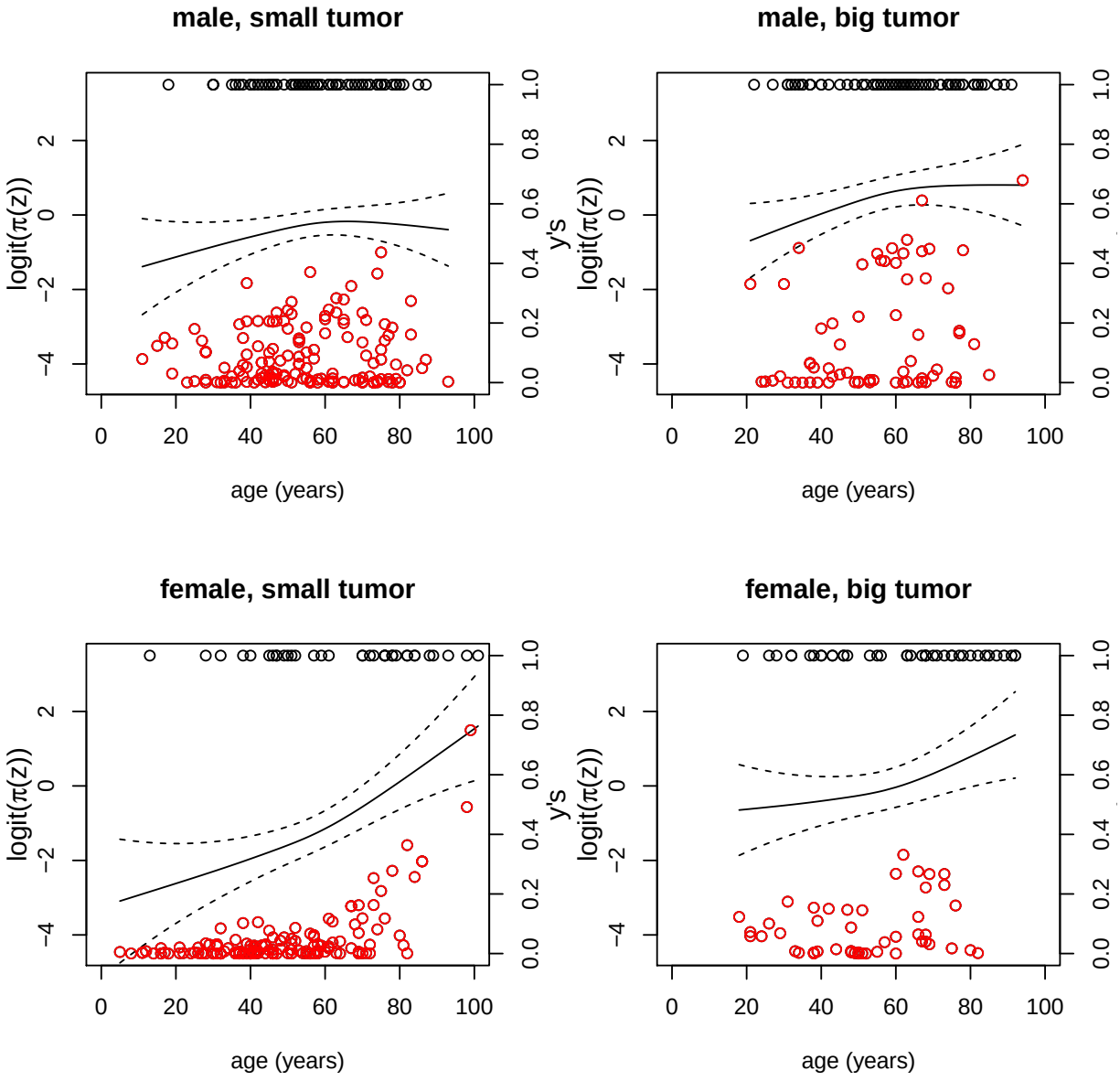

Fig. 2.4. Estimated logit cure rates and their confidence intervals against age. The first row is for Male; the second row is for Female. The left column is for Small tumor size and the right column is for Big tumor size. Superimposed are true data points with positions determined by age and converged $y's$. Black circles are observed failures, red circles are observed censoring.

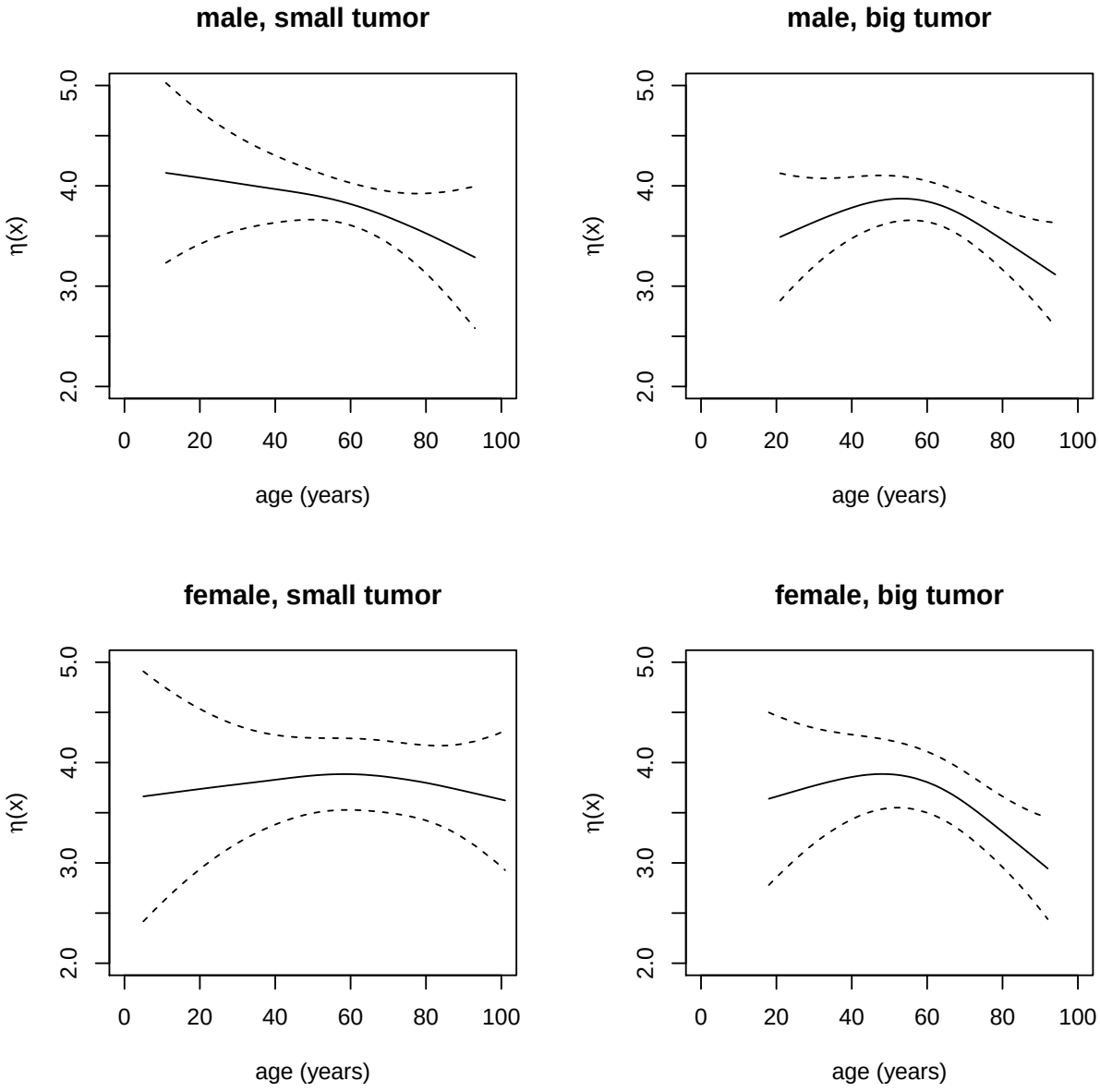

Fig. 2.5. Estimated $\eta(\boldsymbol{x})$ functions and their confidence intervals against age. The first row is for Male; the second row is for Female. The left column is for Small tumor size and the right column is for Big tumor size.

2.7 Proof of the identifiability of model (2.4)

Proof: We need to show that $S_{pop}(t, \boldsymbol{u}) = S^*_{pop}(t, \boldsymbol{u})$ if and only if $(\pi(\boldsymbol{u}), \nu, \eta(\boldsymbol{u})) = (\pi^*(\boldsymbol{u}), \nu^*, \eta^*(\boldsymbol{u}))$. It is obvious that the "if" part is true. To show the "only if" part, suppose that $S_{pop}(t, \boldsymbol{u}) = S^*_{pop}(t, \boldsymbol{u})$. Then we have

$$\frac{\pi(\boldsymbol{u})}{\pi^*(\boldsymbol{u})} = \frac{1 - \exp(-(\frac{t}{e^{\eta^*(\boldsymbol{u})}})^{\nu^*})}{1 - \exp(-(\frac{t}{e^{\eta(\boldsymbol{u})}})^{\nu})} := c(\boldsymbol{u}) \tag{2.17}$$

Define $r_0(\boldsymbol{u}) := \frac{e^{-\eta(\boldsymbol{u})}}{e^{-\eta(\mathbf{0})}}$, $r^*_0(\boldsymbol{u}) := \frac{e^{-\eta^*(\boldsymbol{u})}}{e^{-\eta^*(\mathbf{0})}}$, $S_0(t, \nu) := \exp(-(te^{-\eta(\mathbf{0})})^{\nu})$, and $S^*_0(t, \nu^*) := \exp(-(te^{-\eta^*(\mathbf{0})})^{\nu^*})$.

Now the second equation in (2.17) becomes

$$\big(S^*_0(t, \nu^*)\big)^{r^*_0(\boldsymbol{u})^{\nu^*}} = 1 - c(\boldsymbol{u}) + c(\boldsymbol{u})\big(S_0(t, \nu)\big)^{r_0(\boldsymbol{u})^{\nu}},$$

which is equivalent to

$$r^*_0(\boldsymbol{u})^{\nu^*} = \frac{\log\big(1 - c(\boldsymbol{u}) + c(\boldsymbol{u})S_0(t, \nu)^{r_0(\boldsymbol{u})^{\nu}}\big)}{\log S^*_0(t, \nu^*)}. \tag{2.18}$$

Let $y = c(\boldsymbol{u})$, then $r(\boldsymbol{u}) = r(c^{-1}(y))$. (2.18) becomes

$$r^*_0(c^{-1}(y))^{\nu^*} = \frac{\log\big(1 - y + yS_0(t, \nu)^{r_0(c^{-1}(y))^{\nu}}\big)}{\log S^*_0(t, \nu^*)}. \tag{2.19}$$

For the right hand side of (2.19), the mean value theorem gives

$$\begin{aligned}
&\frac{\log\left(1-y+yS_0(t,\nu)^{r_0(c^{-1}(y))^\nu}\right)}{\log S_0^*(t,\nu^*)}\\
&=\frac{\log\left(1-c(\mathbf{0})+c(\mathbf{0})S_0(t,\nu)^{r_0(\mathbf{0})^\nu}\right)}{\log S_0^*(t,\nu^*)}+(c(\boldsymbol{u})-c(\mathbf{0}))\Big(\frac{d}{dy}\frac{\log\left(1-y+yS_0(t,\nu)^{r_0(c^{-1}(y))^\nu}\right)}{\log S_0^*(t,\nu^*)}\Big)\Big|_{y_0}\\
&=1+(c(\boldsymbol{u})-c(\mathbf{0}))\Big(\frac{d}{dy}\frac{\log\left(1-y+yS_0(t,\nu)^{r_0(c^{-1}(y))^\nu}\right)}{\log S_0^*(t,\nu^*)}\Big)\Big|_{y_0},
\end{aligned}$$

where y_0 is between $c(0)$ and $c(\boldsymbol{u})$. The left hand side of (2.19) not being a function of t requires the right hand side not being a function of t, which leads to $c(\boldsymbol{u})-c(0)=0$. Hence $y=c(\boldsymbol{u})$ is constant. (2.17) then gives $(\nu^*-\nu)\log t=\nu^*\eta^*(\boldsymbol{u})-\nu\eta(\boldsymbol{u})$ for any t. Thus $\nu=\nu^*$ and $\eta(\boldsymbol{u})=\eta^*(\boldsymbol{u})$, and $S_{pop}(t,\boldsymbol{u})$ is uniquely represented and identifiable.

2.8 Extension to Other AFT Distributions and Model Comparison

In model (2.2), Weibull distribution is used in the AFT model mainly due to its simplicity in the survival function form and its flexibility in interpretation. The mixture model can be easily extended to other distributions in the AFT family such as log-logistic, log-normal, gamma, and inverse Gaussian distributions. In this section, log-normal and log-logistic distributions are used for the non-cured subpopulation and the model comparison with Weibull distributions through the likelihood on testing data is performed with a 5-fold cross validation.

First we check the simulation performance of the mixture model with non-cured part assuming a log-normal or log-logistic distribution. For the failure time T with a log-normal distribution, T can be written as $T = \log(X)$, where $X = \sigma Z + \eta$ and Z is a standard normal random variable. Denoting $\mu = e^{\eta}$ and $\nu = 1/\sigma$, the survival function can be written as

$$S(t|\boldsymbol{x}) = 1 - \Phi(\log(\frac{t}{e^{\eta(\boldsymbol{x})}})^{\nu}), \tag{2.20}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution, $e^{\eta(x)}$ is the counterpart of the scale parameter and ν is the counterpart of the shape parameter in the Weibull distribution. If failure time T has a log-logistic distribution, the survival function can be written as

$$S(t|\boldsymbol{x}) = (1 + (\frac{t}{e^{\eta(\boldsymbol{x})}})^{\nu})^{-1}, \tag{2.21}$$

where $e^{\eta(x)}$ is the counterpart of the scale parameter and ν is the counterpart of the shape parameter in the Weibull distribution. For the simulations we use the same functions for π, η and ν as in Section 2.5, where $\pi_2(z) = -0.0278 + 0.7\sin(2(z + 0.6))$, $e^{\eta_1(x)} = .3 \times [10^6 x^{11}(1 - x)^6 + 10^4 x^3(1 - x)^{10}] + 1$ and ν is fixed as 2. So the failure time for non-cured subpopulation are randomly generated according to the survival function (2.20) and the censoring time are generated using an independent Weibull distribution. The simulation results in this specific setting with sample size $n = 800$ are illustrated in Figure (2.6), where it's easy to see the performance of estimation and inference are similar to the ones in the Weibull scenario as expected.

Then the mixture model with log-normal distribution is applied to the melanoma study. The estimated ζ, η and ν are shown is Figure (2.7) and Figure (2.8). With the melanoma data we tried three different AFT distributions (Weibull, log-normal and log-logistic) in the hazard part and they provide similar performance. The logit function of the non-cure rate ζ and the converged latent variables Z_i are almost the same in each setting (the results of log-logistic distribution is a bit different than the other two where the estimation of ζ is a bit alleviated), and the trend of the scale parameter η are similar. For model comparison, an intuitive way is to compare the likelihoods of the fitted models. However in our case the smoothing spline is found through optimizing the penalized log likelihood function. So looking at the likelihood directly may encourage the more curved model with better fit but poor prediction ability. So the model comparison is done using likelihood in a K-fold cross validation way. We split the data into K roughly equal-sized parts; for our scenario $K = 5$. For the kth part, we fit the model to the other $K - 1$ parts of the data, and calculate the negative of log-likelihood of the fitted model when predicting the kth part of the data (predicting ζ, η, and ν). We do this for k = 1, 2, . . .,K and combine the K negative log-likelihood. The following table lists the average of the negative of log-likelihood for the prediction part. Among the three models the one with the Weibull distribution of the hazard part has the smallest negative log-likelihood, indicating the best model among these three for the melanoma data is the one with Weibull distribution assumption in terms of the K-fold criterion.

Model	Average Negative Log-likelihood
Weibull	192.66
Log-logistic	282.20
Log-normal	227.48

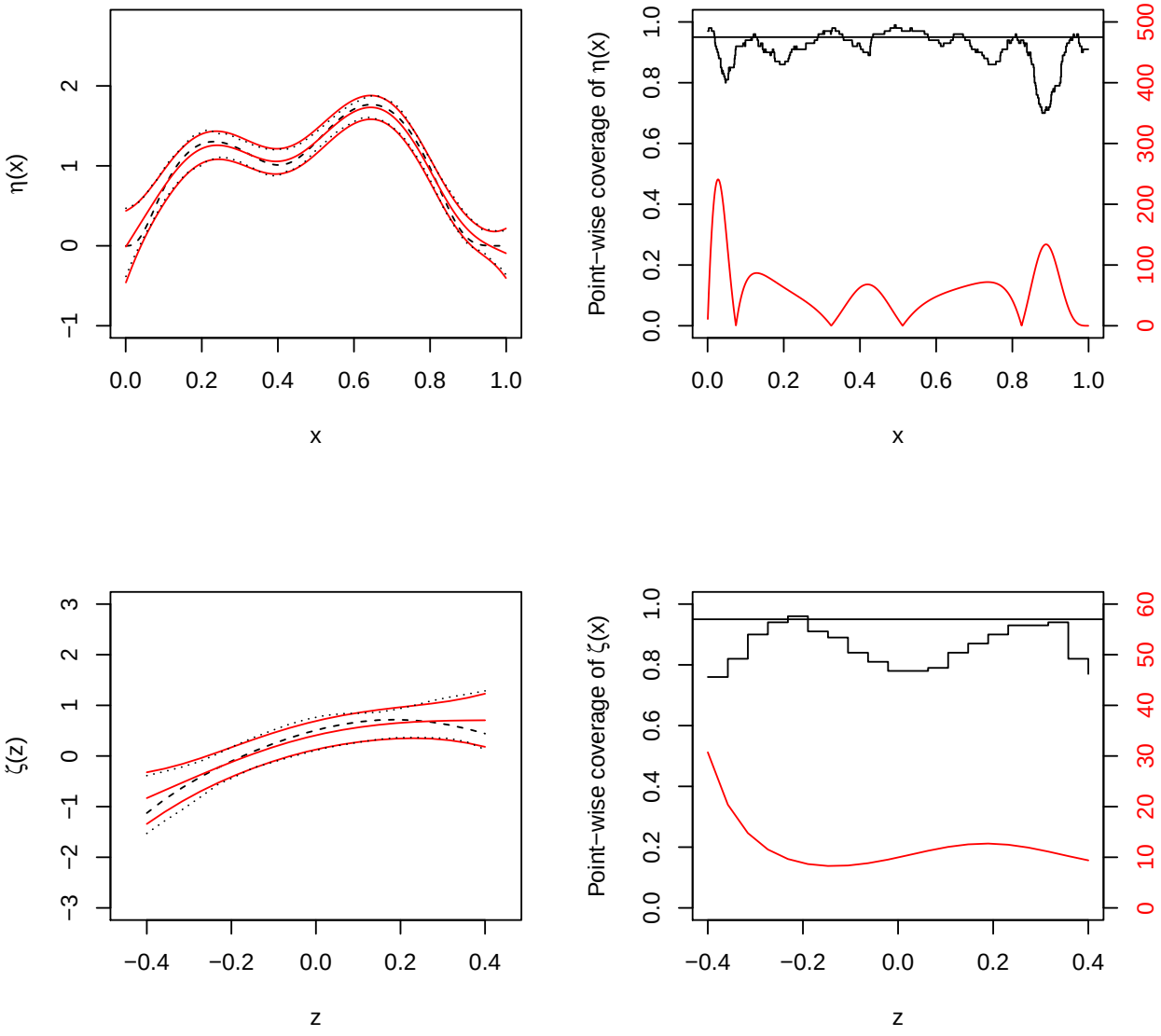

Fig. 2.6. The non-cured part in the mixture model assumes Log-normal distribution. Simulation Results for Test Functions $\eta_1(\boldsymbol{x})$, $\zeta_2(\boldsymbol{z})$ and $n = 800$. Right column: Point-wise coverage (top black lines). Superimposed are nominal coverage (black straight lines) and scaled $|\eta''(x)|$ or $|\zeta''(z)|$. Left column: True functions (dashed) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (solid) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates. The average of the estimated ν is 2.05 (the true $\nu = 2$) and the average of the 95% CIs is $[1.86, 2.23]$. The empirical 2.5 and 97.5 percentiles of the estimated ν is 1.88 and 2.24. The coverage of ν is 0.95.

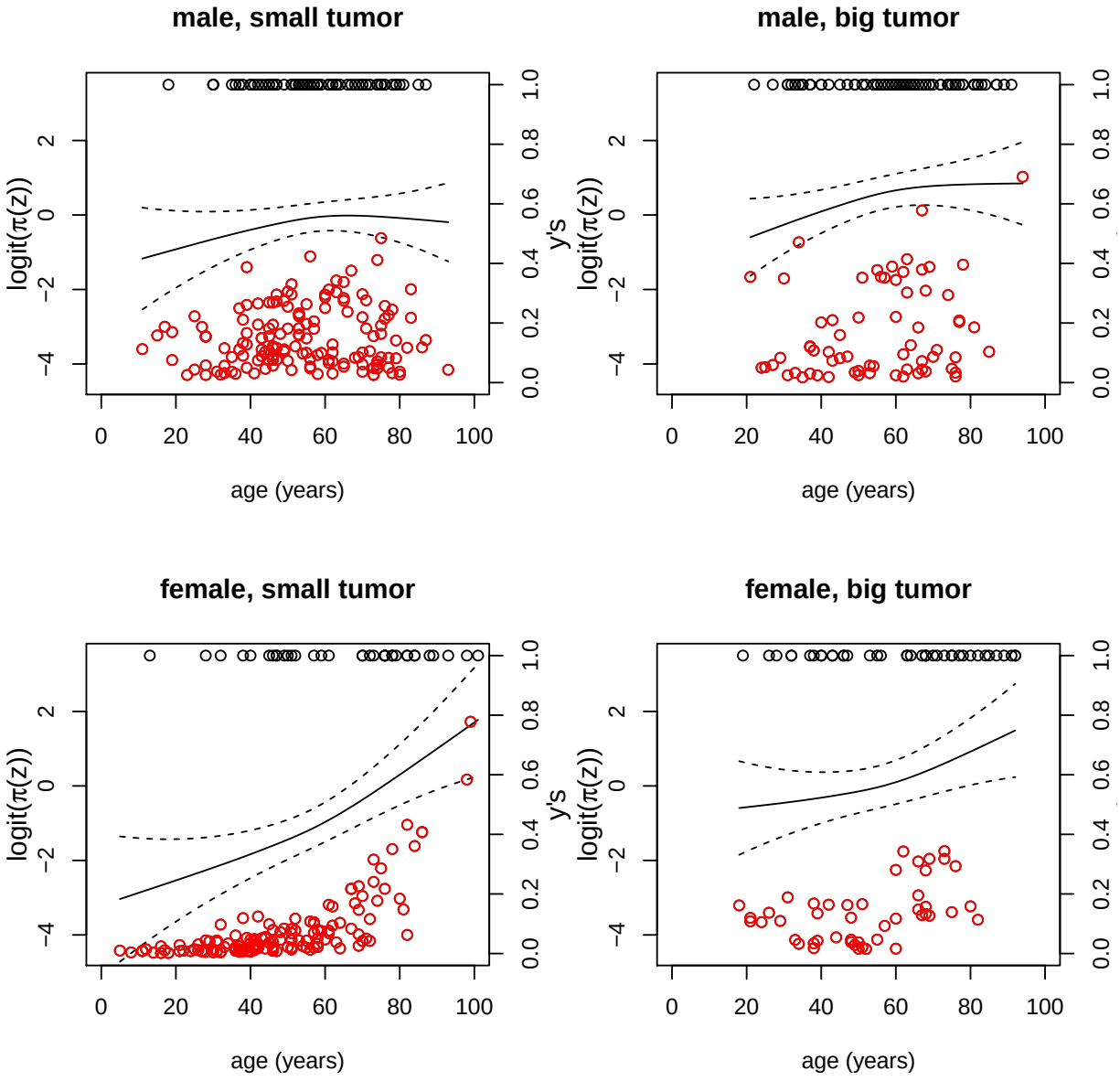

Fig. 2.7. Assume Log-normal distribution in hazard part. Estimated logit cure rates and their confidence intervals against age. The first row is for Male; the second row is for Female. The left column is for Small tumor size and the right column is for Big tumor size. Superimposed are true data points with positions determined by age and converged $y's$. Black circles are observed failures, red circles are observed censoring.

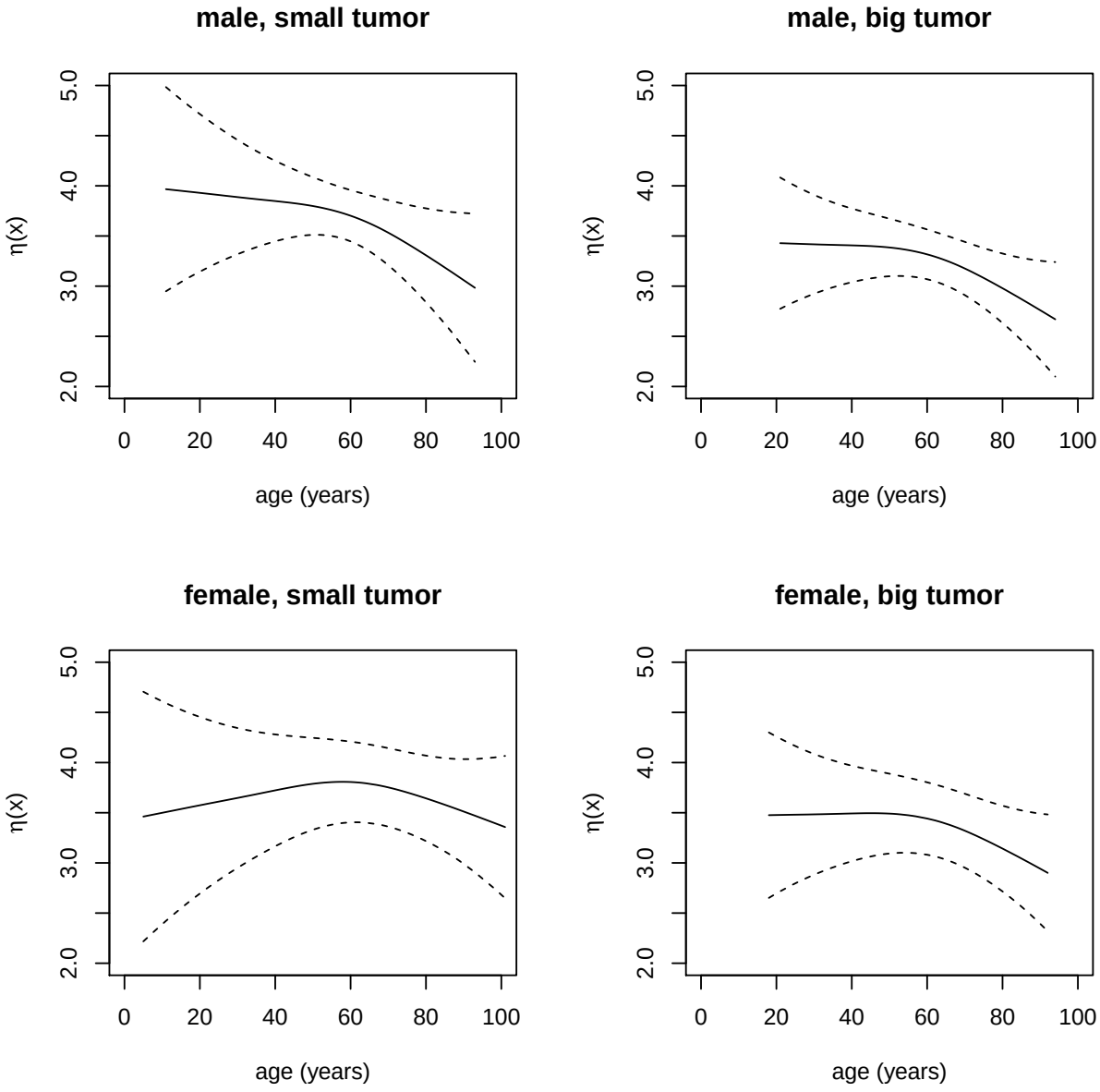

Fig. 2.8. Assume Log-normal distribution in hazard part. Estimated $\eta(\boldsymbol{x})$ functions and their confidence intervals against age. The first row is for Male; the second row is for Female. The left column is for Small tumor size and the right column is for Big tumor size. The estimated ν is 1.10 with confidence interval $[0.97, 1.22]$.

3. PROMOTION TIME CURE MODEL WITH NONPARAMETRIC SPLINE ESTIMATED COMPONENTS

3.1 Introduction

Although the two component mixture model is a popular approach, it has some drawbacks from both a Bayesian and frequentist perspective, as pointed out by Chen et al. [1999] and Ibrahim et al. [2001]. For example, when including covariates through the parameter π via a standard binomial regression model, (2.1) yields improper posterior distributions for many types of noninformative improper priors, including the uniform prior for the regression coefficients. This is a crucial drawback of (2.1) because it implies that Bayesian inference with (2.1) essentially requires a proper prior. In addition, (2.1) does not appear to describe the underlying biological process generating the failure time, at least in the context of cancer relapse, where cure rate models are frequently used. An alternative but equally general model, namely promotion cure model has been proposed in which the survival distribution for cured subjects and non-cured subjects are integrated into one improper survival time distribution, see, e.g., Yakovlev and Tsodikov [1996], Tsodikov [1998], Chen et al. [1999],

Tsodikov et al. [2003], Zeng et al. [2006]. In these models, the population survival function is assumed to have the form

$$S_{\text{pop}}(t, \boldsymbol{x}) = \exp[-\theta(\boldsymbol{x})F(t)], \tag{3.1}$$

where F is an unknown distribution of a nonnegative random variable and $\theta > 0$ can be modeled as $\theta(\boldsymbol{x})$ to incorporate the covariate $\boldsymbol{x}$. This type of model first appeared in Yakovlev and Tsodikov [1996] and Tsodikov [1998] who also noted that the model provided a natural way to extend the proportional hazards regression model. Chen et al. [1999] gave a biological interpretation of the form (3.1) and proposed a Bayesian analysis method with $\theta(x) = \exp(\boldsymbol{x}^T\beta)$ for some unknown parameter vector β. They also pointed out a mathematical connection between (3.1) and the two-component mixture cure models. Zeng et al. [2006] considered a more general form of model $S_{\text{pop}}(t|\boldsymbol{x}) = G_\gamma(\theta(\boldsymbol{x})F(t))$, where G_γ, as γ varies, offers more transformation. This type of model has some distinct advantages such as providing a biologically meaningful interpretation of the model result and allowing construction of a rich class of nonlinear transformation regression models to describe complex covariate effects. In Zeng et al. [2006], they give a biological interpretation to the model (3.1). For the ith subject, let N_i denote the number of tumor cells that have the potential of metastasizing, that is, the number of metastasis-competent tumor cells. The N_i's are unobservable latent variables. We assume that N_i has a Poisson distribution with Poisson rate (mean) $\theta(\boldsymbol{x}_i)$. We denote the promotion time for the kth tumor cell by $\tilde{T}_k$ $(k = 1, \cdots, N_i)$, which is the time for the kth metastasis-competent tumor cell to produce a detectable tumor mass. The $\tilde{T}_k$'s

are also unobservable quantities. Conditional on N_i, the $\tilde{T}_k$'s are independent and identically distributed (iid) as F, where F is sometimes referred to as the promotion time cumulative distribution function. Then the time to relapse of cancer, defined as $T = \min(\tilde{T}_1, \cdots, \tilde{T}_k)$, which is the observed event time, has the survival function $S(t, \boldsymbol{x}_i) = \exp[-\theta(\boldsymbol{x}_i)F(t)]$. One critical assumption for this model is that, conditional on the number of tumor cells, $N_i = k$, $(\tilde{T}_1, \cdots, \tilde{T}_k)$ are mutually independent. This assumption may be unrealistic, because $(\tilde{T}_1, \cdots, \tilde{T}_k)$ are unobserved random variables taken on the same subject. After introducing a subject-specific frailty as a relaxation and remedy of this assumption, a more general form of the survival function for the whole population is derived as

$$S(t, \boldsymbol{x}) = G[\theta(\boldsymbol{x})F(t)], \tag{3.2}$$

where $\theta(\cdot, \cdot)$ is a known link function indexed by unknown parameters β, $F(t)$ is an unspecified distribution function, and G is a known transformation function. Note that the first model is only a special case of (3.2) when $G(\cdot) = \exp(-(\cdot))$. When t is finite, the model gives a semiparametric form for the survival function of susceptible subjects. When t is infinite, the proportion of non-susceptible subjects is given by $G[\theta(x, \beta)]$. For example, when G is $\exp(-(\cdot))$ and θ is linear in β, the promotion cure model assumes a proportional hazards model with parametric covariate effect, and the cure rate is simply $\exp(-\theta(x, \beta))$.

The existing promotion time cure models have a common limitation in that they all model covariate effects in a parametric form whose validity is generally not justified in practice.

The strict parametric assumption can be particularly problematic at the exploratory stage of a study. This calls for more flexible nonparametric modeling of covariate effects. To avoid the model misspecification in a parametric analysis, we assume that in model (3.2), $\theta(\boldsymbol{x}) = \eta(\zeta(\boldsymbol{x}))$, where $\eta(\cdot)$ is a known and strictly positive link function and $\zeta(\cdot) \in \mathcal{C}^{(2)}$ is allowed to take a nonparametric form. An example of $\eta(\cdot)$ is $\exp(\cdot)$ in Zeng and Lin [2007] and the model (3.2) reduces to the usual linear transformation models studied by Cheng, Wei, and Ying [1995].

The likelihood function of the parameters (ζ, F) is given by

$$\prod_{i=1}^{n} \Big\{ \left[\{-G' \left(\eta(\zeta(\boldsymbol{x}_i))F(X_i)\right) \eta(\zeta(\boldsymbol{x}_i)) f(X_i) \}^{\delta_i} \{ G \left(\eta(\zeta(\boldsymbol{x}_i))F(X_i)\right) \}^{(1-\delta_i)} \right]^{I(X_i<\infty)} \times \left[G \left(\eta(\zeta(\boldsymbol{x}_i))\right) \right]^{I(X_i=\infty)} \Big\}, \tag{3.3}$$

where $G''(x)$ denotes the second derivative of G with respect to x.

We wish to maximize the foregoing likelihood function to obtain the maximum likelihood estimators (MLEs) ζ and F; however, this maximum does not exist, because one can choose $f(X_i) = \infty$ for some X_i with $\delta_i = 1$. Thus we apply a nonparametric maximum likelihood estimation approach, where F is allowed to be a right-continuous function. Instead of maximizing (3.3), we maximize the following modified function,

$$\prod_{i=1}^{n} \Big\{ \left[\{-G' \left(\eta(\zeta(\boldsymbol{x}_i))F(X_i)\right) \eta(\zeta(\boldsymbol{x}_i)) F\{X_i\} \}^{\delta_i} \{ G \left(\eta(\zeta(\boldsymbol{x}_i))F(X_i)\right) \}^{(1-\delta_i)} \right]^{I(X_i<\infty)} \times \left[G \left(\eta(\zeta(\boldsymbol{x}_i))\right) \right]^{I(X_i=\infty)} \Big\}, \tag{3.4}$$

where $F\{X_i\}$ is the jump size of F at X_i.

The MLE for F is termed the nonparametric maximum likelihood estimator (NPMLE) for F, and it is easy to show that the estimate for F must be a distribution function only with point masses at the observed X_i with $\delta_i = 1$. To estimate $F(t)$ nonparametrically, we must determine a follow-up time such that all censored observations beyond that follow-up time, called the cure threshold, are treated as $X_i = \infty$ (i.e., observed to be cured) and all observations lower than this threshold are treated as $X_i < \infty$ (i.e., observed to be either a failure or right-censored). This assumption is needed so that the model is identifiable in (ζ, F). Note that if a parametric form is assumed for F (as in Ibrahim et al. 2001), then the condition that some of the X_i's are observed to be infinity is not needed.

3.2 Computation of Profile Likelihood

It is too difficult to compute the MLEs for F and ζ from (3.4) directly. We follow a two-stage profile likelihood approach and first assume that ζ is fixed and find the F that maximize (3.4), which is equivalent to maximize

$$\sum_{i=1}^{n} I(X_i < \infty) \left[\delta_i \log p_i + \delta_i \log \left\{-G'(\eta(\zeta(\boldsymbol{x}_i))F_i)\right\} + (1-\delta_i) \log \left\{G(\eta(\zeta(\boldsymbol{x}_i))F_i)\right\}\right], \quad (3.5)$$

where $p_i = F\{X_i\}$ and $F_i = \sum_{X_j \le X_i, \delta_j = 1} p_j$ is the cumulative function at X_i and we have the constraint that $\sum_{X_j < \infty, \delta_j = 1} p_j = 1$.

If we order the observed failure times from smallest to largest and use the indices $(1), \cdots, (m)$ for the ordered times, $Y_{(1)} < \cdots < Y_{(m)}$, where $m = \sum_i \delta_i I(Y_i < \infty)$, then, after introducing the Lagrange multiplier λ, we obtain $p_{(i)}$ by solving the equation

$$\begin{aligned} &\frac{1}{p_{(i)}} + \sum_{j=1}^{n} \Big\{ \frac{G''(\eta(\zeta(x_j))F_{(j)})\eta(\zeta(x_j))I(Y_{(i)} \le Y_j < \infty)}{G'(\eta(\zeta(x_i))F_{(i)})} \\ &+ (1-\delta_j)\frac{G'(\eta(\zeta(x_j))F_{(i)})\eta(\zeta(x_j))I(Y_{(i)} \le Y_j < \infty)}{G(\eta(\zeta(x_j))F_{(i)})} \Big\} - \lambda = 0, \end{aligned} \tag{3.6}$$

After some derivations we have the recursive formula

$$\begin{aligned} \frac{1}{p_{(i+1)}} = \frac{1}{p_{(i)}} &+ \sum_{i=1}^{m} \frac{G''(\eta(\zeta(x_{(i)}))F_{(i)})\eta(\zeta(x_{(i)}))}{G'(\eta(\zeta(x_{(i)}))F_{(i)})} \\ &+ \sum_{i=1}^{m} \sum_{Y_{(i)} < Y_j < Y_{(i+1)}} \frac{G'(\eta(\zeta(x_j))F_{(i)})\eta(\zeta(x_j))}{G(\eta(\zeta(x_j))F_{(i)})}, \end{aligned} \tag{3.7}$$

where $F_{(i)} = p_{(1)} + \cdots + p_{(i)}$. Let $p_{(m)} = \alpha$ and using the fact that $\sum p_{(m)} = 1$ we can calculate $p_{(i)}$ from $p_{(i+1)}$ as a function of α. Thus the likelihood (3.5) can be determined only by α and ζ. Hence to obtain an estimator of (F, ζ) we maximise the profile likelihood with respect to (α, ζ).

3.3 Computation of Penalized Likelihood

From the last section we can treat ζ, α as independent parameters and $p_{(1)}, \ldots, p_{(m-1)}$ as functions of ζ and α, and we want to minimize $-\log L(\zeta, \alpha) + \lambda_0 J(\zeta)$ under the constraint $\sum_{i=1}^{m} p_{(i)} = 1$, which is equivalent to minimize

$$-\log L(\zeta, \alpha) + \lambda_0 J(\zeta) + \lambda_1 (\sum_{i=1}^{m} p_{(i)} - 1), \tag{3.8}$$

with λ_1 performing as the Lagrange multiplier. The first term of (3.8) depends on ζ only through the evaluations $[x_i]\zeta = \zeta(x_i)$, so the argument of Section 2.3.2 in Gu [2002] applies and the minimizer ζ_λ of (3.8) has an expression

$$\begin{aligned} \zeta(x) &= \sum_{i=1}^{m} d_{\nu,\zeta}\phi_{\nu,\zeta}(x) + \sum_{i=1}^{n} c_{i,\zeta} R_{J,\zeta}(x_i, x) = \boldsymbol{\phi}_\zeta^T(x)\boldsymbol{d}_\zeta + \boldsymbol{\xi}_\zeta^T(x)\boldsymbol{c}_\zeta \\ &= \begin{pmatrix} \boldsymbol{\phi}_\zeta(x) \\ \boldsymbol{\xi}_\zeta(x) \end{pmatrix}^T \begin{pmatrix} \boldsymbol{d}_\zeta \\ \boldsymbol{c}_\zeta \end{pmatrix} \stackrel{Let}{=} \boldsymbol{\psi}_\zeta(x)^T \boldsymbol{b}_\zeta \end{aligned} \tag{3.9}$$

Then the constrained maximum likelihood equations for $\boldsymbol{c}$, $\boldsymbol{d}$ and $p_{(1)}, \ldots, p_{(m)}$ can be reduced to solving the following score equations for $\boldsymbol{c}$, $\boldsymbol{d}$, α, and λ_1,

$$\begin{cases} \dfrac{\partial}{\partial \boldsymbol{c}}(-\log L(\zeta,\alpha) + \lambda_0 c^T Q c + \lambda_1(\sum_{i=1}^{m} p_{(i)} - 1)) = 0 \\ \dfrac{\partial}{\partial \boldsymbol{d}}(-\log L(\zeta,\alpha) + \lambda_0 c^T Q c + \lambda_1(\sum_{i=1}^{m} p_{(i)} - 1)) = 0 \\ \dfrac{\partial}{\partial \alpha}(-\log L(\zeta,\alpha) + \lambda_0 c^T Q c + \lambda_1(\sum_{i=1}^{m} p_{(i)} - 1)) = 0 \\ \sum_{i=1}^{m} p_{(i)} - 1 = 0 \end{cases}$$

If we choose the strictly positive link function η simply as $\eta(x) = e^x$, these equations are

$$\begin{aligned} &\sum_{i=1}^{m} \frac{1}{p_{(i)}} \frac{\partial}{\partial \boldsymbol{c}} p_{(i)} + \sum_{i=1}^{m} \frac{G''(e^{\zeta(x_i)} F_{(i)})}{G'(e^{\zeta(x_i)} F_{(i)})} \{ \frac{\partial}{\partial \boldsymbol{c}} e^{\zeta(x_i)} F_{(i)} + e^{\zeta(x_i)} \frac{\partial}{\partial \boldsymbol{c}} F_i \} \\ &+ \sum_{i=1}^{m} \sum_{Y_{(i)} < Y_j < Y_{(i+1)}} \frac{G'(e^{\zeta(x_j)} F_{(i)})}{G(e^{\zeta(x_j)} F_{(i)})} \{ \frac{\partial}{\partial \boldsymbol{c}} e^{\zeta(x_j)} F_{(i)} + e^{\zeta(x_j)} \frac{\partial}{\partial c} F_i \} \\ &+ \sum_{j=1}^{n} \delta_j \frac{\partial}{\partial \boldsymbol{c}} \zeta(x_j) + \sum_{j=1}^{n} I(Y_j = \infty) \frac{G'(e^{\zeta(x_j)})}{G(e^{\zeta(x_j)})} \frac{\partial}{\partial \boldsymbol{c}} e^{\zeta(x_j)} - 2\lambda_0 Q \boldsymbol{c} - \lambda_1 \sum_{i=1}^{m} \frac{\partial}{\partial \boldsymbol{c}} p_{(i)} = 0 \end{aligned} \tag{3.10a}$$

$$\begin{aligned} &\sum_{i=1}^{m} \frac{1}{p_{(i)}} \frac{\partial}{\partial \boldsymbol{d}} p_{(i)} + \sum_{i=1}^{m} \frac{G''(e^{\zeta(x_i)} F_{(i)})}{G'(e^{\zeta(x_i)} F_{(i)})} \{ \frac{\partial}{\partial \boldsymbol{d}} e^{\zeta(x_i)} F_{(i)} + e^{\zeta(x_i)} \frac{\partial}{\partial \boldsymbol{d}} F_i \} \\ &+ \sum_{i=1}^{m} \sum_{Y_{(i)} < Y_j < Y_{(i+1)}} \frac{G'(e^{\zeta(x_j)} F_{(i)})}{G(e^{\zeta(x_j)} F_{(i)})} \{ \frac{\partial}{\partial \boldsymbol{d}} e^{\zeta(x_j)} F_{(i)} + e^{\zeta(x_j)} \frac{\partial}{\partial \boldsymbol{d}} F_i \} \\ &+ \sum_{j=1}^{n} \delta_j \frac{\partial}{\partial \boldsymbol{d}} \zeta(x_j) + \sum_{j=1}^{n} I(Y_j = \infty) \frac{G'(e^{\zeta(x_j)})}{G(e^{\zeta(x_j)})} \frac{\partial}{\partial \boldsymbol{d}} e^{\zeta(x_j)} - \lambda_1 \sum_{i=1}^{m} \frac{\partial}{\partial \boldsymbol{d}} p_{(i)} = 0 \end{aligned} \tag{3.10b}$$

$$\sum_{i=1}^{m}\frac{1}{p_{(i)}}\frac{\partial}{\partial\alpha}p_{(i)}+\sum_{i=1}^{m}\frac{G''(e^{\zeta(x_i)}F_{(i)})}{G'(e^{\zeta(x_i)}F_{(i)})}e^{\zeta(x_i)}\frac{\partial}{\partial\alpha}F_i$$
$$+\sum_{i=1}^{m}\sum_{Y_{(i)}<Y_j<Y_{(i+1)}}\frac{G'(e^{\zeta(x_j)}F_{(i)})}{G(e^{\zeta(x_j)}F_{(i)})}e^{\zeta(x_j)}\frac{\partial}{\partial\alpha}F_i-\lambda_1\sum_{i=1}^{m}\frac{\partial}{\partial\alpha}p_{(i)}=0 \tag{3.10c}$$

$$\sum_{i=1}^{m}p_{(i)}-1=0 \tag{3.10d}$$

After eliminating λ_1 from the first three equations, the Newton-Raphson algorithm can be used to solve the system of equations in (3.10a)-(3.10d). The first and second derivatives of $p_{(i)}$ with respect to $\boldsymbol{c}$, $\boldsymbol{d}$ and α can be computed using the recursive formula (3.7).

When we try to find the optimized penalty parameter λ_0, the cross-validation method using Kullback-Leibler distance does not work well as in Gu [2002] because $p_{(i)}$s make the situation more complicated in this application. So we resort to the standard K-fold cross validation to find the optimized penalty parameter λ_0.

3.4 Inference

We now derive the confidence intervals for the ζ function. An adequately justified interval estimate is a rarity in nonparametric function estimation. Wahba (1978, 1983) derived Bayes confidence intervals from the Bayes model, which is equivalent to the penalized likelihood estimation in a reproducing kernel Hilbert space $\mathcal{H}$ with the penalty $J(f)$ a square (semi) norm.

For the Gaussian-type responses penalized likelihood

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i-\eta(x_i))^2+\lambda\sum_{\beta=1}^{p}\theta^{-1}(\eta,\eta)_\beta, \tag{3.11}$$

a smoothing spline minimizing this likelihood is a Bayes estimate of $\eta=\sum_{\beta=0}^{p}\eta_\beta$, where η_0 has a diffuse prior in $\mathcal{H}_0$ and η_β, $\beta=1,\cdots,p$, have mean zero Gaussian process priors on $\mathcal{X}$ with covariance functions $E[\eta_\beta(x)\eta_\beta(y)]=b\theta_\beta R_\beta(x,y)$, independent of each other, where $b=\sigma^2/n\lambda$, see Gu (2002).

For the general penalized likelihood $L(\eta)+\frac{\lambda}{2}J(\eta)$ ($L(\eta)$ is the minus log likelihood), when a quadratic penalty is used, the $\frac{\lambda}{2}J(\eta)$ part can be viewed as the log density for certain Gaussian prior on the coefficient vector of η. So the penalized likelihood becomes the log posterior distribution for η. Hence the minimizer $\hat{\eta}$ is the posterior mode, whose asymptotic variance can be derived via a quadratic approximation of the penalized likelihood. This approximation essentially approximates the posterior distribution of η by a normal distribution at the minimizer $\hat{\eta}$. The confidence intervals derived in this manner for various problems (Gaussian regression, non-Gaussian regression, hazard estimation, etc.) can be found in Wahba (1983) and Gu (1992).

The approach discussed above cannot be easily used in our case because we use the penalized profile likelihood in which there's a nuisance parameter $p_{(m)}=\alpha$. Instead we estimate the asymptotic variance of using the negative inverse of the curvature of the penalized profile likelihood. We use the parametrization of $\zeta=\phi^{\mathbf{T}}\boldsymbol{d}+\xi^{\mathbf{T}}\boldsymbol{c}$ and refer the function

$\zeta = \phi^{\mathbf{T}}\boldsymbol{d} + \xi^{\mathbf{T}}\boldsymbol{c} = \psi^{\mathbf{T}}\boldsymbol{b}$ and the coefficients $(\boldsymbol{d}^T, \boldsymbol{c}^T)^T = \boldsymbol{b}$ interchangeably. And the penalized profile likelihood is $pl(\boldsymbol{b}, \alpha) = -\log L(\zeta, \alpha) + \lambda_0 J(\zeta)$. Then the asymptotic variance of $(\hat{\mathbf{b}}, \hat{\alpha})$ can be estimated using the negative inverse of the curvature of the penalized profile likelihood $pl(\boldsymbol{b}, \alpha)$ at at $(\hat{\boldsymbol{b}}, \hat{\alpha})$, which is

$$H := -\left.\begin{pmatrix} \frac{\partial^2}{\partial \boldsymbol{b}^2} pl(\boldsymbol{b}, \alpha) & \frac{\partial^2}{\partial \boldsymbol{b} \partial \alpha} pl(\boldsymbol{b}, \alpha) \\ \frac{\partial^2}{\partial \alpha \partial \boldsymbol{b}} pl(\boldsymbol{b}, \alpha) & \frac{\partial^2}{\partial \alpha^2} pl(\boldsymbol{b}, \alpha) \end{pmatrix}\right|_{\boldsymbol{b}=\hat{\boldsymbol{b}}, \alpha=\hat{\alpha}}.$$

Denote $\varphi^T = (\psi(\boldsymbol{x})^T, 0)$. Then the approximate variance for $\hat{\zeta}(\boldsymbol{x}) = (\hat{\boldsymbol{b}}, \hat{\alpha})^T \varphi$ is $s(\boldsymbol{x})^2 = \varphi^T H \varphi$.

3.5 Simulation

A simulation study is carried out to study the performance of the proposed promotion cure model with spline estimated components. If we choose $G(x) = \exp(-x)$, our model can be written as

$$S(t|\boldsymbol{x}) = \exp(-e^{\zeta(\boldsymbol{x})} F(t)),$$

where $\zeta(\boldsymbol{x})$ and $F(t)$ are the nonparametric functions we want to estimate. For the ζ function, we tried two univariate test functions:

$$\zeta_1(x) = .1 \times [10^6 x^{11}(1-x)^6 + 10^4 x^3 (1-x)^{10}]$$

$$\zeta_2(x) = 9.81^{1.4} * 3.5 * (x)^{2.5}/400$$

For the F function, we use $F(t) = 1 - \exp(-t)$ (exponential distribution with rate parameter=1). For each setting of testing functions, we repeat with sample size $n = 400$ and $n = 800$. So in total we have 4 simulations. We choose a censoring function such that the overall censoring rate is about 50%. One hundred replicates were generated for each setting. The point-wise 95% confidence intervals were calculated for $\zeta(x)$ on a x grid of size 200 equally spaced on $[0, 1]$. The coverage results in all simulation settings are very similar.

Figure 3.1 shows simulation results for test function $\zeta_1(x)$ and sample size $n = 800$. Plotted in the left frame are the true test functions (black solid), the averages of point-wise function estimates (red dashed), the averages of point-wise 95% CIs (red dashed), and the empirical 2.5% and 97.5% percentiles of point-wise function estimates (grey dotted). The frame on the right shows the point-wise coverage of the test function against covariate. Figure 3.3 shows the same simulation results for test function $\zeta_2(x)$ and sample size $n = 800$.

On the right columns of these graphs, it's easy to see that the point-wise coverage is close to the nominal level 0.95. And on the left columns of the graphs the means of the estimated confidence interval are close to empirical 2.5% and 97.5% percentile of the 100 function estimates (The bands are connected point-wise intervals, with no simultaneous coverage property intended.) The widths of the intervals appear to be of the proper magnitude. We also noticed that the width of the point-wise confidence intervals is increasing towards the ends of the axis where information from the data is diminishing. There are a couple

of factors that would negatively affect the coverage. First, similar to the observations of Wahba [1983], Nychka [1988], and Gu [1992] in regression settings of smoothing spline, lower coverage appears to roughly track high curvature. Second, the local sample size was a key factor of coverage property, meaning that if certain area has more data, the estimates of that area are more reliable.

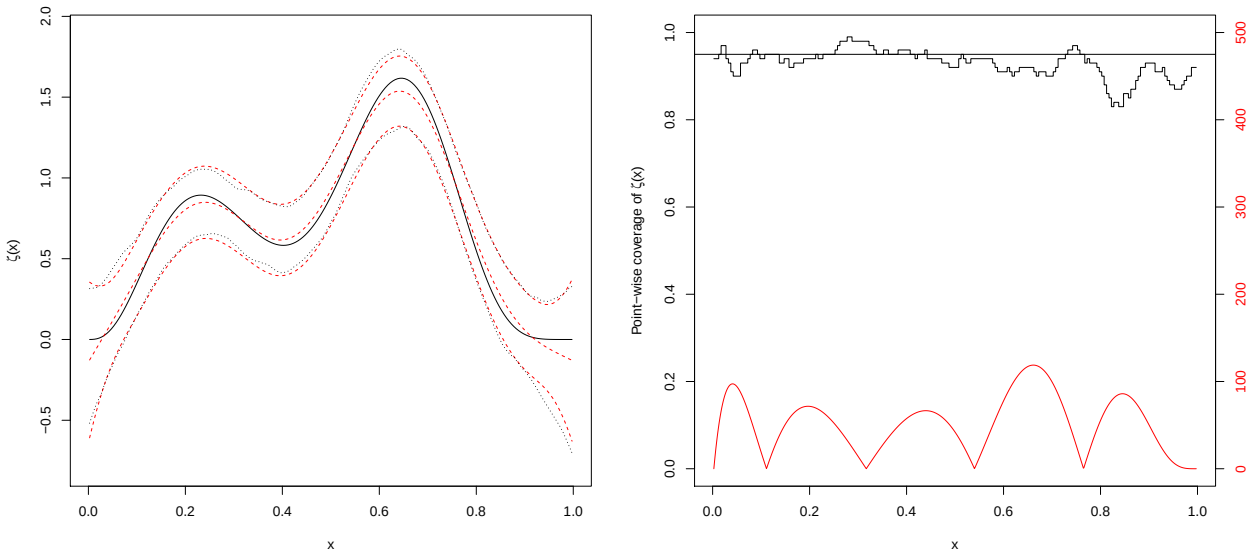

Fig. 3.1. Simulation Results for Test Function $\zeta_1(\boldsymbol{x})$ and $n = 800$. Right graph: Point-wise coverage (top black lines). Superimposed are nominal coverage and scaled $|\zeta''(x)|$. Left graph: True function and the estimate, including averages of point-wise function estimates, averages of point-wise 95% CIs and empirical 2.5 and 97.5 percentiles of point-wise function estimates, all based on 100 data replicates.

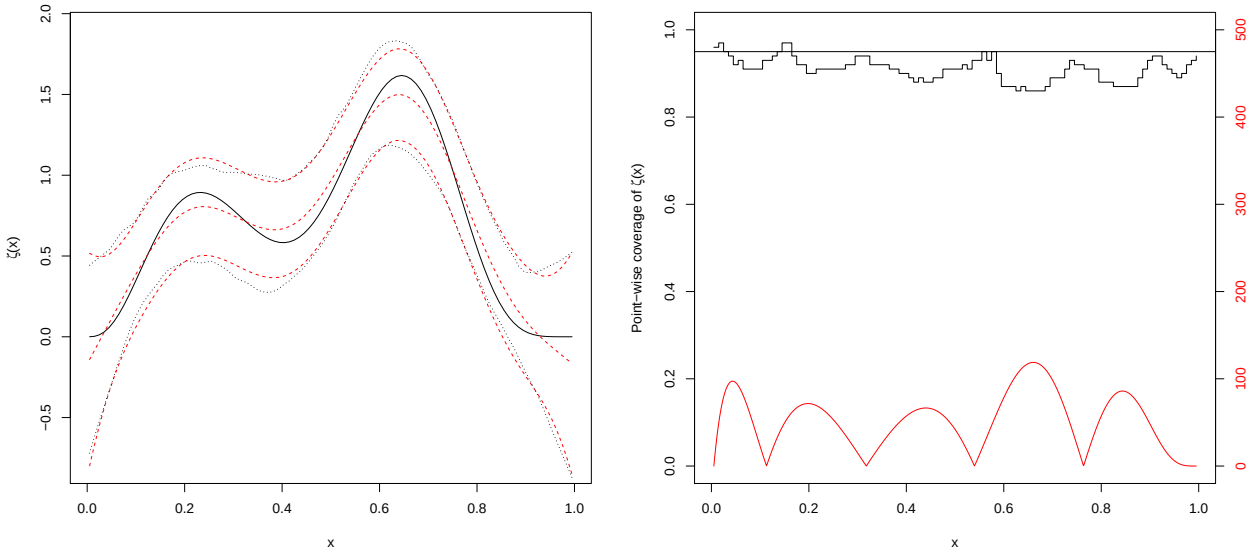

Fig. 3.2. Simulation Results for Test Function $\zeta_1(\boldsymbol{x})$ and $n = 400$. Right graph: Point-wise coverage (top black lines). Superimposed are nominal coverage and scaled $|\zeta''(x)|$. Left graph: True function and the estimate, including averages of point-wise function estimates, averages of point-wise 95% CIs and empirical 2.5 and 97.5 percentiles of point-wise function estimates, all based on 100 data replicates.

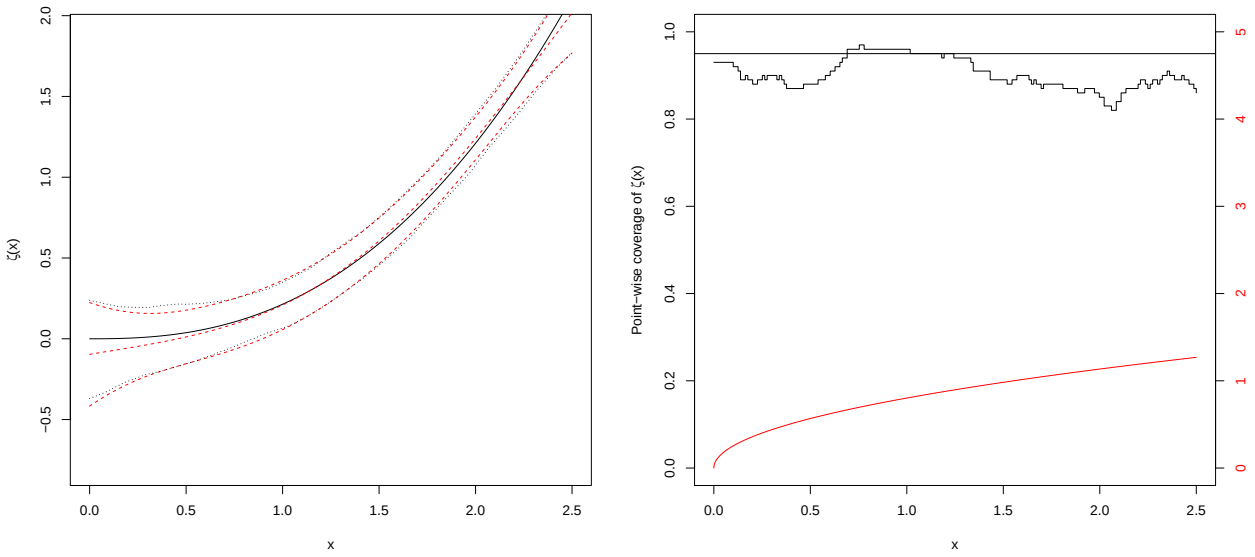

Fig. 3.3. Simulation Results for Test Function $\zeta_2(\boldsymbol{x})$ and $n = 800$. Right graph: Point-wise coverage (top black lines). Superimposed are nominal coverage and scaled $|\zeta''(x)|$. Left graph: True function and the estimate, including averages of point-wise function estimates, averages of point-wise 95% CIs and empirical 2.5 and 97.5 percentiles of point-wise function estimates, all based on 100 data replicates.

3.6 Melanoma Cancer Data

Wang et al. [2012] fitted a nonparametric mixture model to a dataset from a melanoma cancer study, in which 637 patients' information (gender, age, tumor size) were recorded as covariates. Time to relapse or death was used as the endpoint. The overall censoring rate is 63.3% with a large proportion of long-time censoring events, which might be an evidence that a cure rate model is appropriate in this situation. We applied the proposed promotion

time cure model approach to analyze the dataset. The covariates $\boldsymbol{x}$ is (age, gender, size). We allow interaction between (age, gender, size) for $\zeta(\boldsymbol{x})$ such that

$$\begin{aligned}\zeta(age, gender, size) = \zeta_0 &+ \zeta_a(age) + \zeta_g(gender) + \zeta_s(size) \\ &+ \zeta_{ag}(age, gender) + \zeta_{as}(age, size) + \zeta_{gs}(gender, size) \\ &+ \zeta_{ags}(age, gender, size)\end{aligned}$$

where $\zeta_g(gender)$ and $\zeta_s(size)$ only have the parametric components since *gender* and *size* are both factors. Figure 3.4 and Figure 3.5 respectively plot $\zeta(\boldsymbol{x})$ against age for the four patient groups and the estimated $F(t)$ against failure time for the overall model. As we expected, the ζ function is increasing over *age*, which means that as the patients are aging the survival function is decreasing more quickly and eventually the patient has a lower cure rate. We also notice that Big tumor size corresponds to larger ζ at a given age for both Male and Female groups, which means patients with Big tumor is more like to fail compared with patients with Small tumor. And roughly Male corresponds to larger ζ at a given age for both Big tumor size and Small tumor size groups, which indicates the melanoma cancer in the study is more dangerous to males than females.

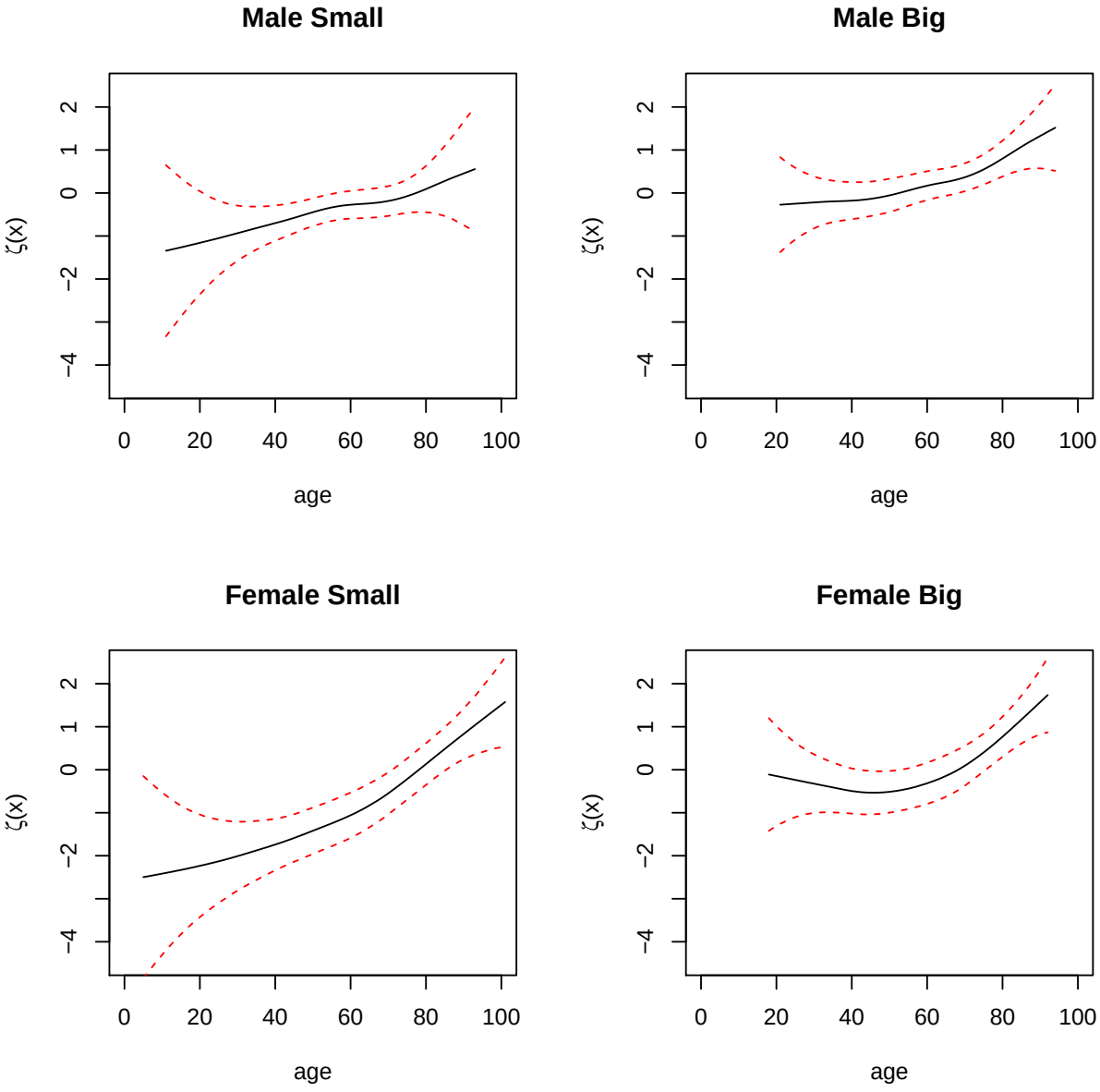

Fig. 3.4. Estimated function $\zeta(\boldsymbol{x})$ and its point-wise confidence intervals (red dashed) when *age* and *tumor size* are fixed. The first row is for Male and the second row is for Female. The left column is for Small tumor size and the right column is for Big tumor size. The penalty parameter is chosen by K-Fold cross validation.

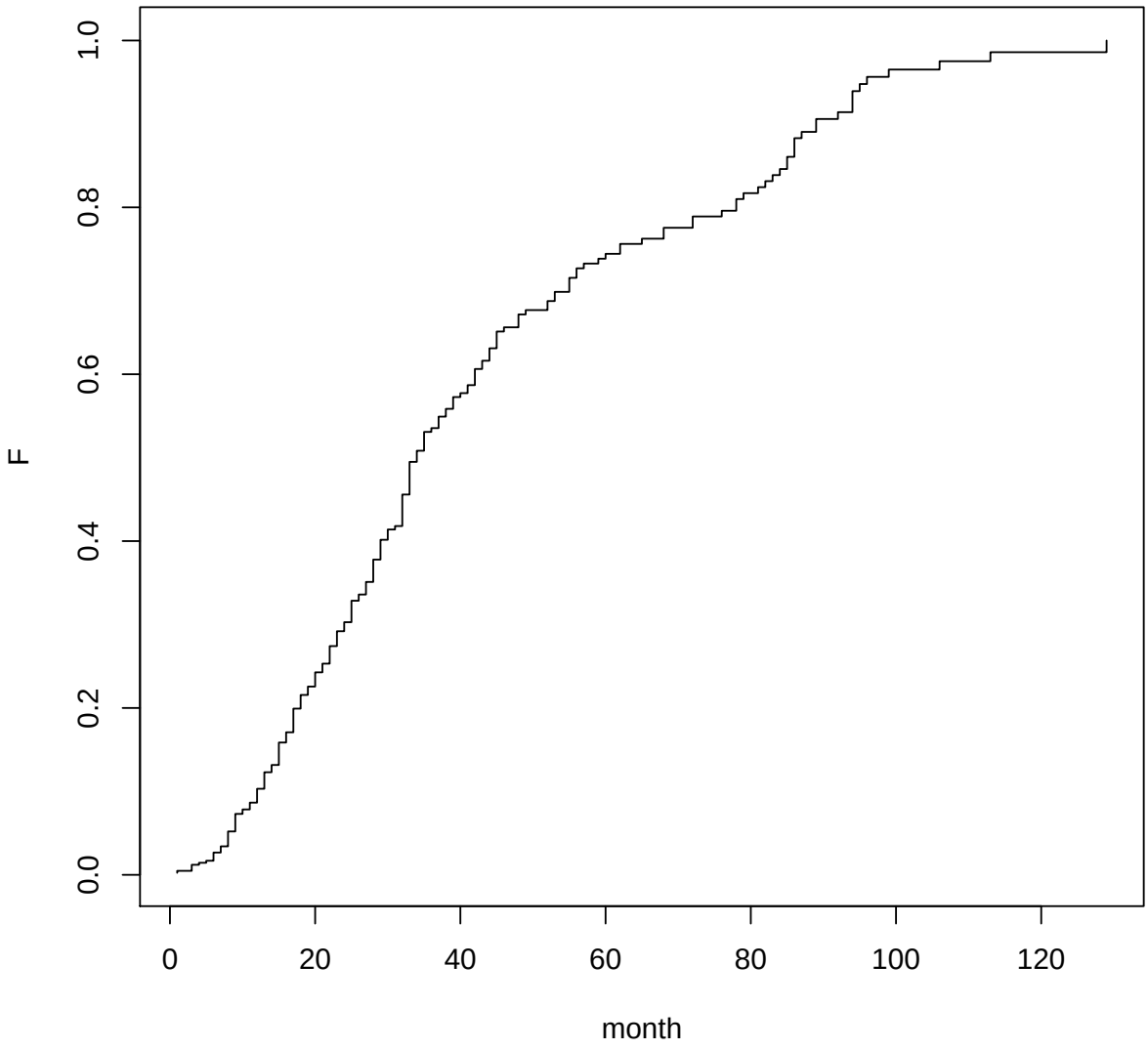

Fig. 3.5. Estimated function $\hat{F}(t)$.

LIST OF REFERENCES

LIST OF REFERENCES

J. Berkson and R. P. Gage. Survival curve for cancer patients following treatment. *J. Amer. Statist. Assoc.*, 47:501–515, 1952.

M.-H. Chen, J. G. Ibrahim, and D. Sinha. A new Bayesian model for survival data with a surviving fraction. *J. Amer. Statist. Assoc.*, 94:909–919, 1999.

Y. Q. Chen and M. Wang. Analysis of accelerated hazards models. *J. Amer. Statist. Assoc.*, 95:608–618, 2000.

S. C. Cheng, L. J. Wie, and Z. Ying. Analysis of transformation models with censored data. *Biometrika*, 82:835–845, 1995.

D. D. Cox and Y. Chang. Iterated state space algorithms and cross validation for generalized smoothing splines. Technical Report 49, Department of Statistics, University of Illinois, Champion, IL, 1990.

D. R. Cox. Regression models and life tables. *J. Roy. Statist. Soc. Ser. B*, 34:187–220 (with discussions), 1972.

D. R. Cox. Some remarks on the analysis of survival data. *the First Seattle Symposium of Biostatistics: Survival Analysis*, 123:1–9, 1997.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 39:1–37 (with discussions), 1977.

B. Efron and D. V. Hinkley. The observed versus expected information. *Biometrika*, 65: 457–487, 1978.

V. Farewell. The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, 38:1041–1046, 1982.

I. J. Good and R. A. Gaskins. Nonparametric roughness penalties for probability densities. *Biometrika*, 58:255–277, 1971.

C. Gu. Adaptive spline smoothing in non Gaussian regression models. *J. Amer. Statist. Assoc.*, 85:801–807, 1990.

C. Gu. Penalized likelihood regression: A Bayesian analysis. *Statist. Sin.*, 2:255–264, 1992.

C. Gu. Penalized likelihood hazard estimation: A general procedure. *Statist. Sin.*, 6: 861–876, 1996.

C. Gu. *Smoothing Spline ANOVA Models*. Springer-Verlag, New York, 2002.

C. Gu. Model diagnostics for smoothing spline ANOVA models. *Canadian J. of Statist.*, 32:347–358, 2004.

C. Gu and C. Qiu. Smoothing spline density estimation: Theory. *Ann. Statist.*, 21:217–234, 1993.

C. Gu and D. Xiang. Cross-validating non-gaussian data: Generalized approximate cross-validation revisited. *J. Comput. Graph. Statist.*, 10(3):581–591, 2001.

Joseph G. Ibrahim, Ming-Hui Chen, and Debajyoti Sinha. *Bayesian Survival Analysis*. Springer-Verlag, 2001.

M. Jamshidian and R. I. Jennrich. Acceleration of the em algorithm by using quasi-newton methods. *J. Roy. Statist. Soc. Ser. B*, 52(2):569587, 1997.

Y.-J. Kim and C. Gu. Smoothing spline Gaussian regression: More scalable computation via efficient approximation. *J. Roy. Statist. Soc. Ser. B*, 66:337–356, 2004.

G. Kimeldorf and G. Wahba. A correspondence between Bayesian estimation of stochastic processes and smoothing by splines. *Ann. Math. Statist.*, 41:495–502, 1970a.

G. Kimeldorf and G. Wahba. Spline functions and stochastic processes. *Sankhya Ser. A*, 32:173–180, 1970b.

G. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *J. Math. Anal. Applic.*, 33:82–85, 1971.

A. Y. C. Kuk and C. H. Chen. A mixture model combining logistic regression with proportional hazards regression. *Biometrika*, 79:531–541, 1992.

C. Liu, D. B. Rubin, and Y. N. Wu. Parameter expansion to accelerate EM: The PX-EM algorithm. *Biometrika*, 85(4):755–770, 1998.

T. A. Louis. Finding the observed information matrix when using the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 44(2):226–233, 1982.

W. Lu and Z. Ying. On semiparametric transformation cure models. *Biometrika*, 91: 331–343, 2004.

X-L. Meng and D. B. Rubin. Maximum likelihood estimation via the ecm algorithm: A general framework. *Biometrika*, 80:267278, 1993.

R. Neal, G. Hintion, and M. I. Jordan. A view of the em algorithm that justifies incremental, sparse, and other variants. *Learning in Graphical Models*, page 355368, 1999.

D. Nychka. Bayesian confidence intervals for smoothing splines. *J. Amer. Statist. Assoc.*, 83:1134–1143, 1988.

M. Othus, Y. Li, and R. C. Tiwari. A class of semiparametric mixture cure survival models with dependent censoring. *J. Amer. Statist. Assoc.*, 104:1241–1250, 2009.

Y. Peng and K. B. G. Dear. A nonparametric mixture model for cure rate estimation. *Biometrics*, 56(1):237–243, 2000.

Nancy Reid. A conversation with sir david cox. *Statistical Science*, 9(3):439–455, 1994.

J. P. Sy and J. M. G. Taylor. Estimation in a Cox proportional hazards cure model. *Biometrics*, 56(1):227–236, 2000.

A. D. Tsodikov. A proportional hazards model taking account of long-term survivors. *Biometrics*, 54:1508–1516, 1998.

A. D. Tsodikov, J. G. Ibrahim, and A. Y. Yakovlev. Estimating cure rates from survival data: An alternative to two-component mixture models. *J. Amer. Statist. Assoc.*, 98: 1063–1078, 2003.

G. Wahba. Bayesian "confidence intervals" for the cross-validated smoothing spline. *J. Roy. Statist. Soc. Ser. B*, 45:133–150, 1983.

G. Wahba. Partial and interaction spline models for the semiparametric estimation of functions of several variables. In *Computer Science and Statistics: Proceedings of the 18th Symposium on the Interface*, pages 75–80, 1986.

G. Wahba. *Spline Models for Observational Data*, volume 59 of CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM, Philadelphia, 1990.

G. Wahba, Y. Wang, C. Gu, R. Klein, and B. E. K. Klein. Smoothing spline ANOVA for exponential families, with application to the Wisconsin Epidemiological Study of Diabetic Retinopathy. *Ann. Statist.*, 23:1865–1895, 1995.

G. Wahba, Y. Lin, and C. Leng. Penalized log likelihood density estimation, via smoothing spline ANOVA and ranGACV. Technical Report 1048, Department of Statistics, University of Wisconsin, Madison, WI, 2001.

L. Wang, P. Du, and Hua Liang. Two-component mixture cure rate model with spline estimated nonparametric components. *Biometrics*, 68(3):726–735, 2012.

M. A. Woodbury. Discussion of paper by Hartley and Hocking. *Biometrics*, 27:808–817, 1971.

D. Xiang and G. Wahba. A generalized approximate cross validation for smoothing splines with non-Gaussian data. *Statist. Sin.*, 6:675–692, 1996.

A. Y. Yakovlev and A. D. Tsodikov. *Stochastic models of tumor latency and their biostatistical applications*. World Scientific, Hackensack, NJ, 1996.

S. Zacks. *The Theory of Statistical Inference*. Wiley, New York, 1971.

D. Zeng and D. Y. Lin. Maximum likelihood estimation in semiparametric regression models with censored data (with discussion). *J. Roy. Statist. Soc. Ser. B*, 69:507–564, 2007.

D. Zeng, G. Yin, and J. G. Ibrahim. Semiparametric transformation models for survival data with a cure fraction. *J. Amer. Statist. Assoc.*, 101(474):670–684, 2006.

J. Zhang and Y. Peng. Accelerated hazards mixture cure model. *Lifetime Data Analysis*, 15:455–467, 2009.

D. M. Zucker and A. F. Karr. Nonparametric survival analysis with time-dependent covariate effects: A penalized partial likelihood approach. *Ann. Statist.*, 18:329–353, 1990.

APPENDIX

A. ADDITIONAL SIMULATIONS

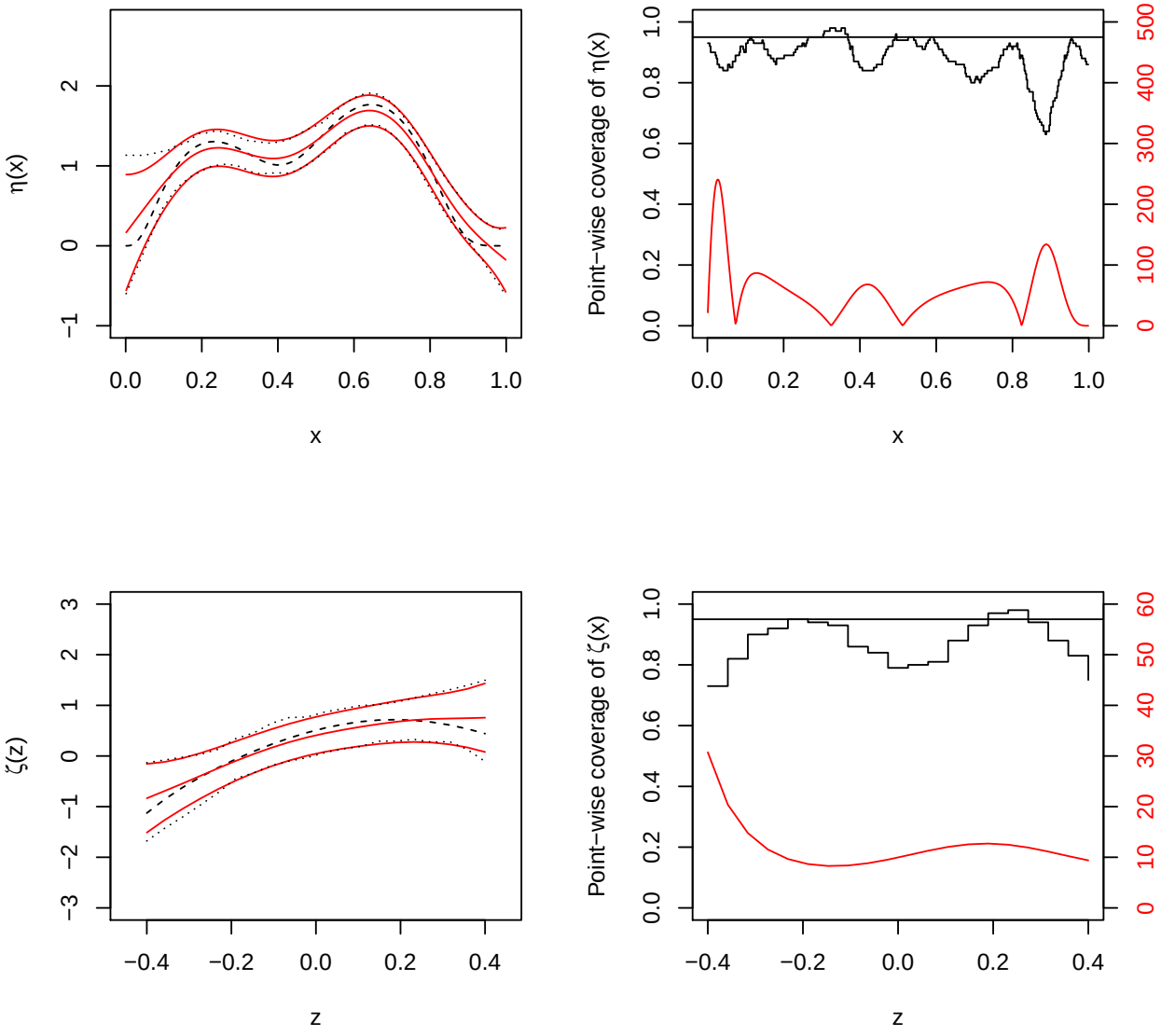

Fig. A.1. Simulation Results for Test Functions $\eta_1(\boldsymbol{x})$, $\zeta_2(\boldsymbol{z})$ and $n = 400$. Right column: Point-wise coverage (top black lines). Superimposed are nominal coverage (black straight lines) and scaled $|\eta''(x)|$ or $|\zeta''(z)|$. Left column: True functions (dashed) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (solid) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates. The average of the estimated ν is 2.05 (the true $\nu = 2$) and the average of the 95% CIs is $[1.79, 2.31]$. The empirical 2.5 and 97.5 percentiles of the estimated ν is 1.83 and 2.29. The coverage of ν is 0.98.

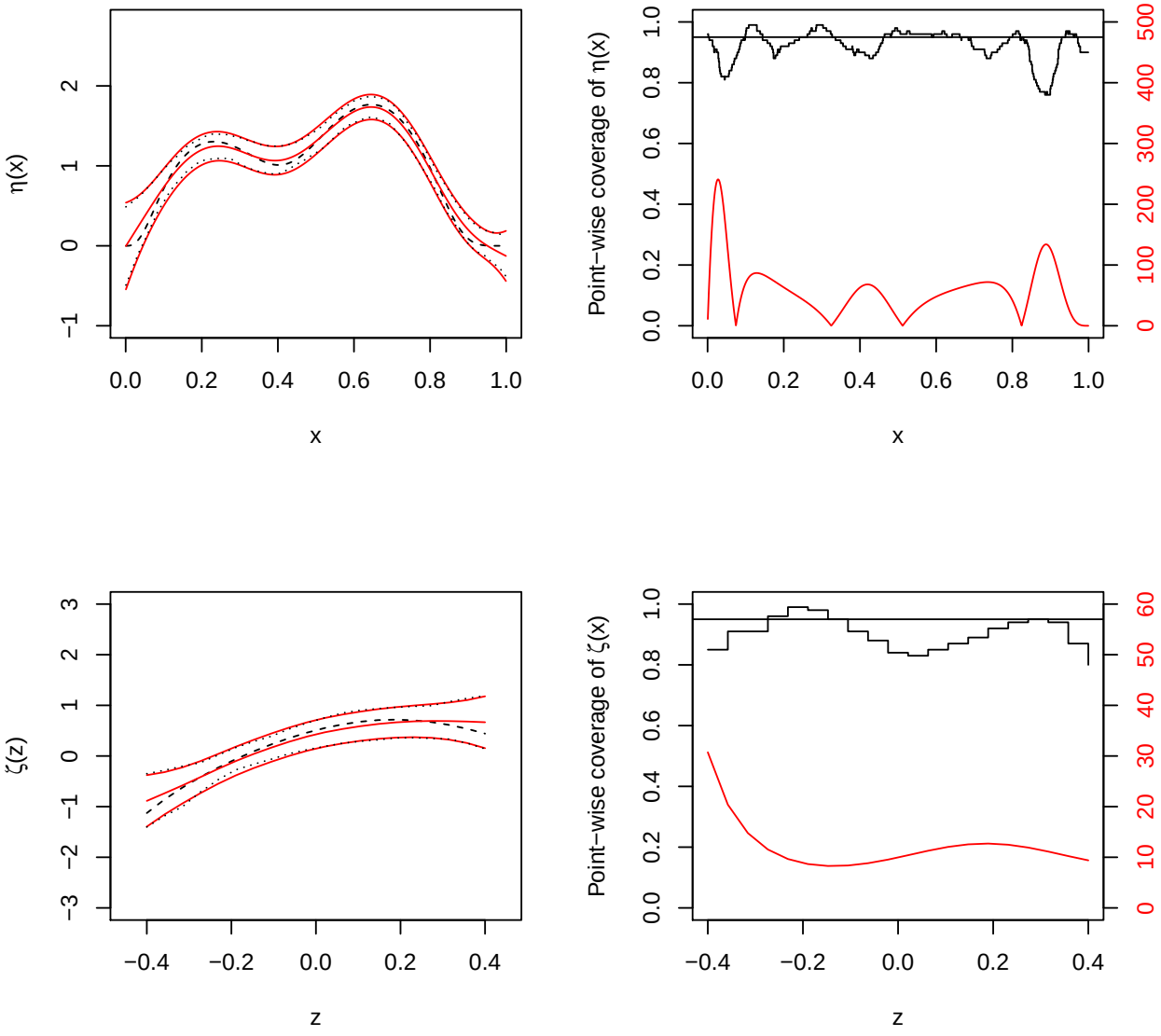

Fig. A.2. Simulation Results for Test Functions $\eta_1(\boldsymbol{x})$, $\zeta_2(\boldsymbol{z})$ and $n = 800$. Right column: Point-wise coverage (top black lines). Superimposed are nominal coverage (black straight lines) and scaled $|\eta''(x)|$ or $|\zeta''(z)|$. Left column: True functions (dashed) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (solid) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates. The average of the estimated ν is 2.02 (the true $\nu = 2$) and the average of the 95% CIs is $[1.83, 2.20]$. The empirical 2.5 and 97.5 percentiles of the estimated ν is 1.85 and 2.21. The coverage of ν is 0.95.

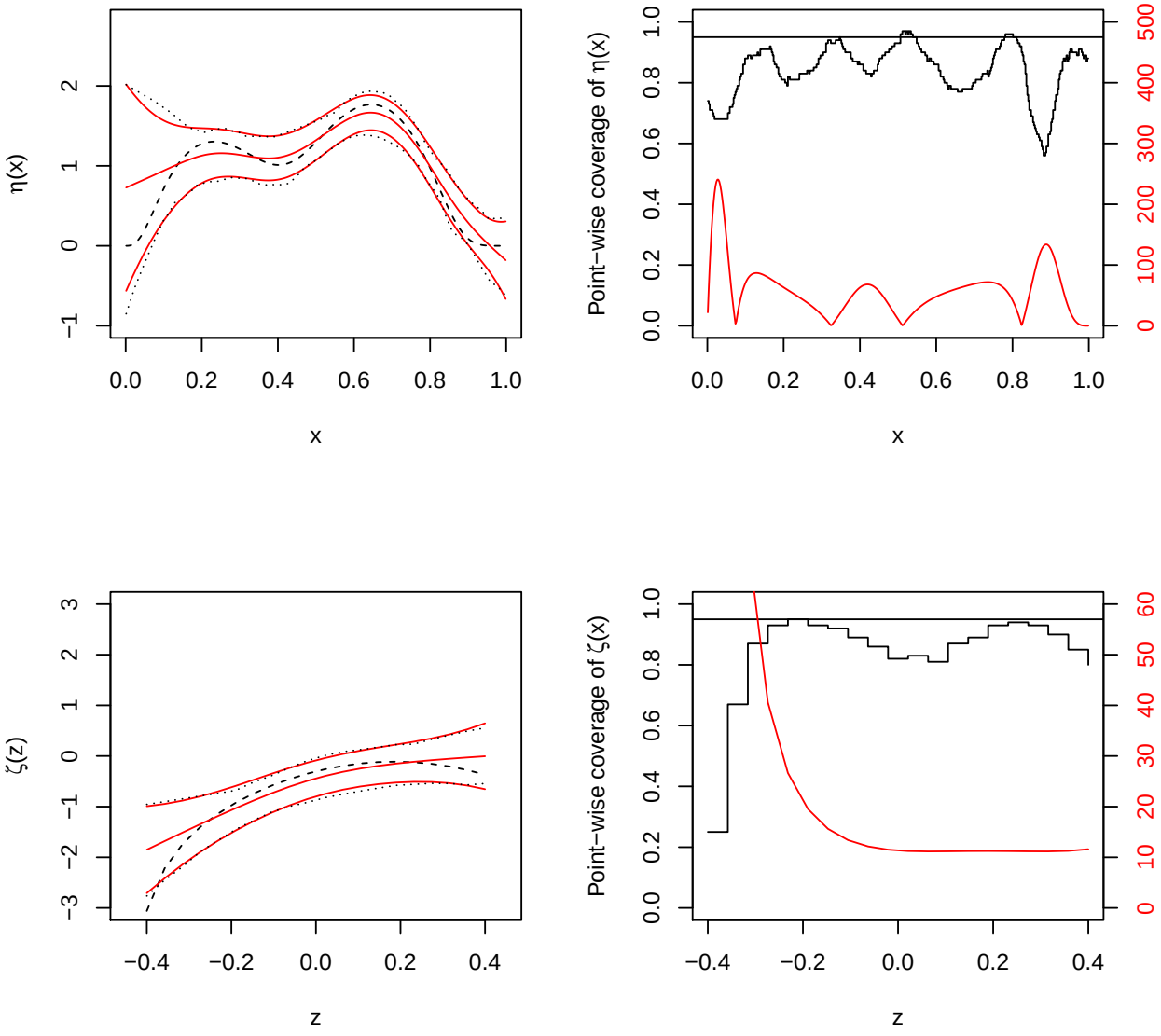

Fig. A.3. Simulation Results for Test Functions $\eta_1(\boldsymbol{x})$, $\zeta_3(\boldsymbol{z})$ and $n = 400$. Right column: Point-wise coverage (top black lines). Superimposed are nominal coverage (black straight lines) and scaled $|\eta''(x)|$ or $|\zeta''(z)|$. Left column: True functions (dashed) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (solid) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates. The average of the estimated ν is 2.01 (the true $\nu = 2$) and the average of the 95% CIs is $[1.72, 2.39]$. The empirical 2.5 and 97.5 percentiles of the estimated ν is 1.79 and 2.52. The coverage of ν is 0.95.

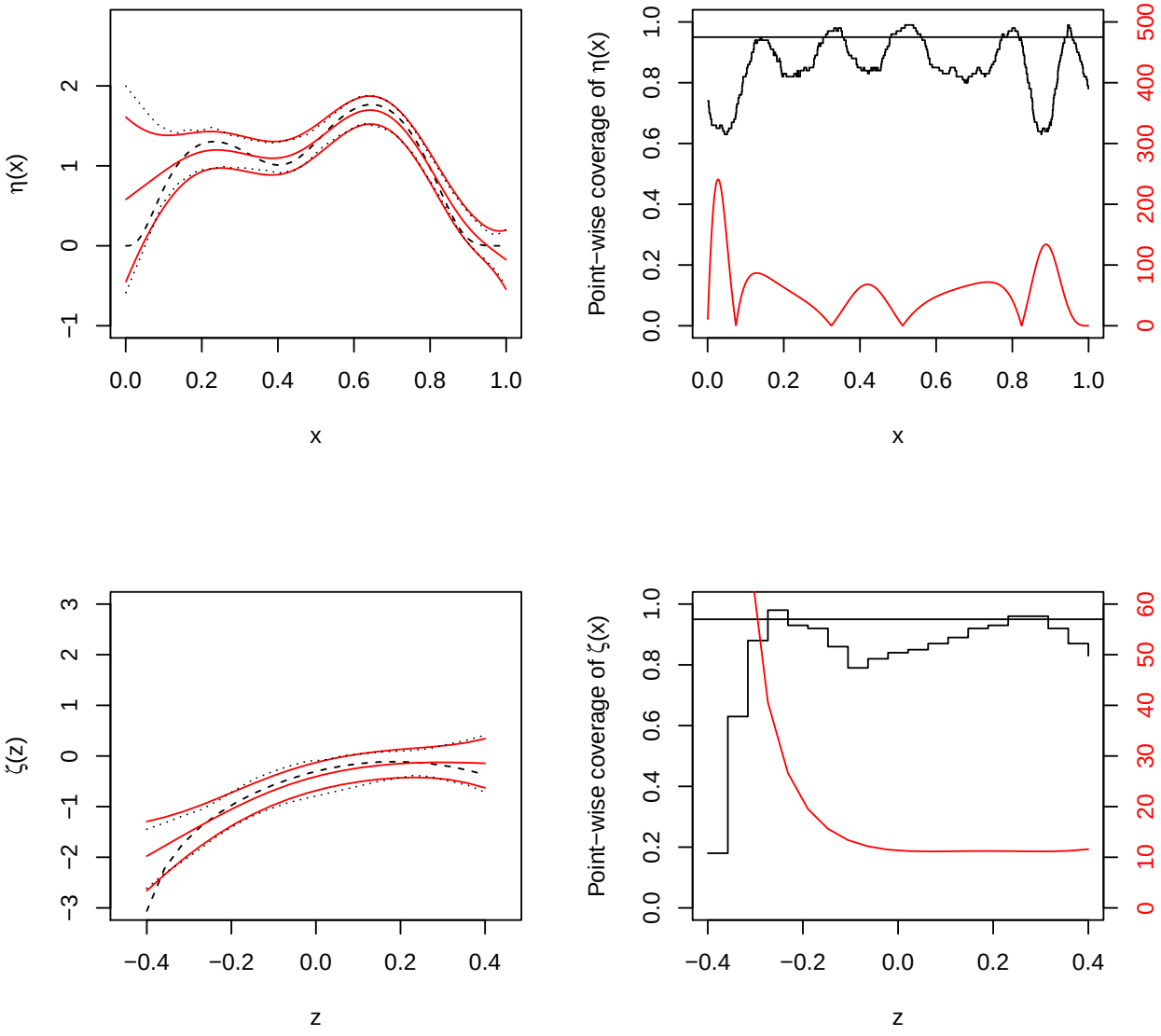

Fig. A.4. Simulation Results for Test Functions $\eta_1(\boldsymbol{x})$, $\zeta_3(\boldsymbol{z})$ and $n = 800$. Right column: Point-wise coverage (top black lines). Superimposed are nominal coverage (black straight lines) and scaled $|\eta''(x)|$ or $|\zeta''(z)|$. Left column: True functions (dashed) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (solid) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates. The average of the estimated ν is 2.03 (the true $\nu = 2$) and the average of the 95% CIs is $[1.80, 2.26]$. The empirical 2.5 and 97.5 percentiles of the estimated ν is 1.86 and 2.25. The coverage of ν is 0.95.

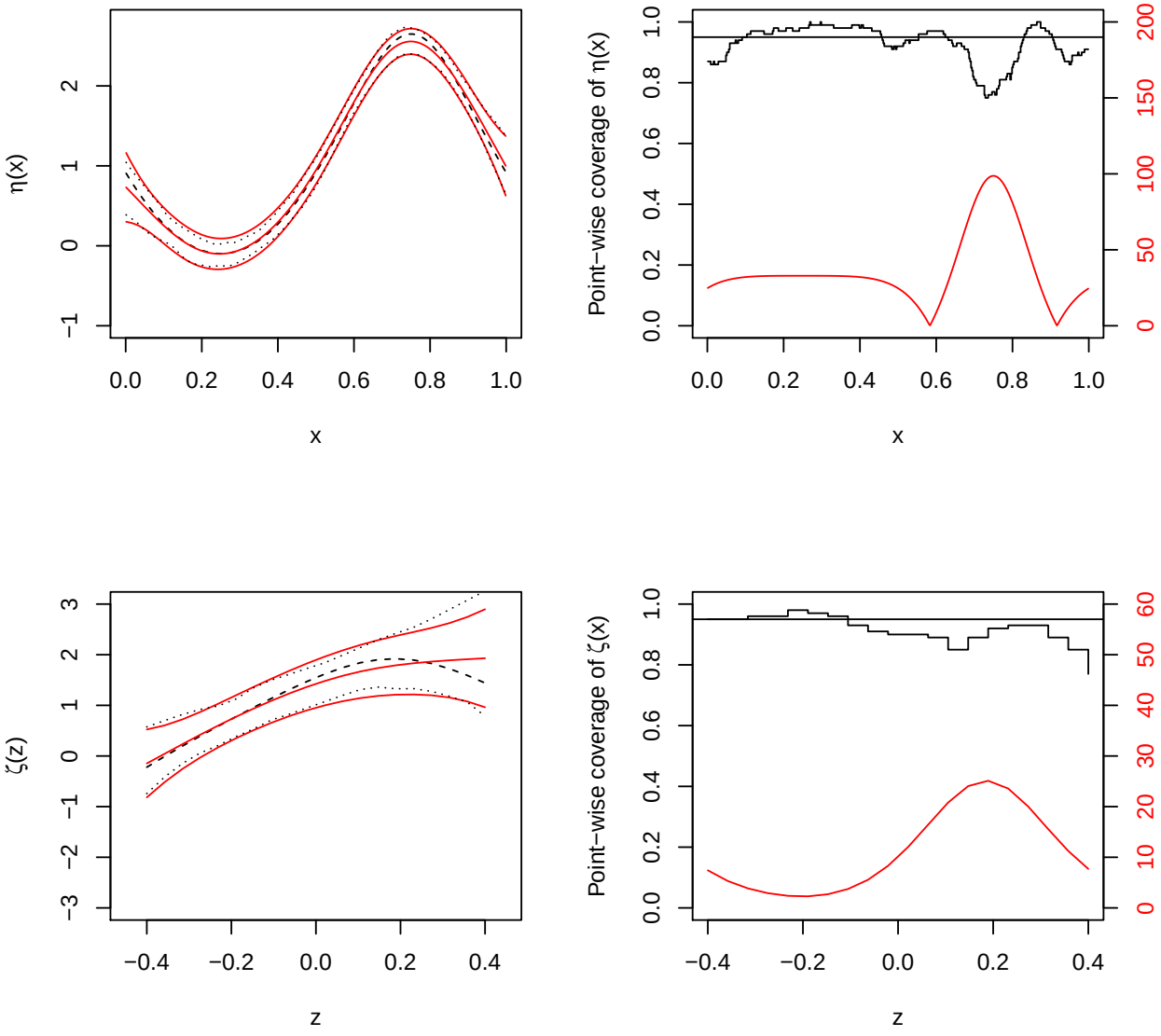

Fig. A.5. Simulation Results for Test Functions $\eta_2(\boldsymbol{x})$, $\zeta_1(\boldsymbol{z})$ and $n = 400$. Right column: Point-wise coverage (top black lines). Superimposed are nominal coverage (black straight lines) and scaled $|\eta''(x)|$ or $|\zeta''(z)|$. Left column: True functions (dashed) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (solid) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates. The average of the estimated ν is 2.06 (the true $\nu = 2$) and the average of the 95% CIs is $[1.83, 2.28]$. The empirical 2.5 and 97.5 percentiles of the estimated ν is 1.89 and 2.32. The coverage of ν is 0.93.

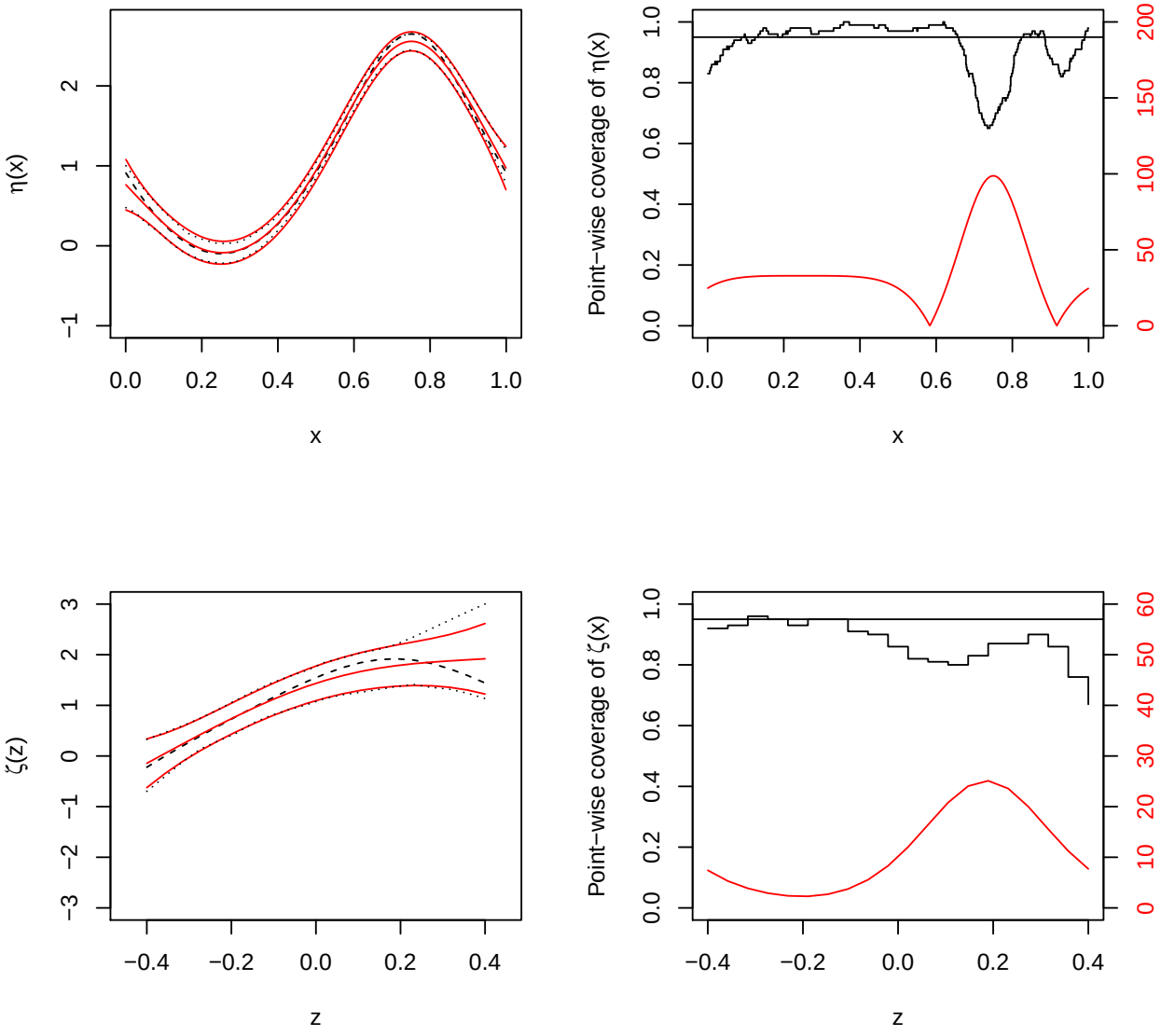

Fig. A.6. Simulation Results for Test Functions $\eta_2(\boldsymbol{x})$, $\zeta_1(\boldsymbol{z})$ and $n = 800$. Right column: Point-wise coverage (top black lines). Superimposed are nominal coverage (black straight lines) and scaled $|\eta''(x)|$ or $|\zeta''(z)|$. Left column: True functions (dashed) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (solid) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates. The average of the estimated ν is 2.03 (the true $\nu = 2$) and the average of the 95% CIs is $[1.87, 2.19]$. The empirical 2.5 and 97.5 percentiles of the estimated ν is 1.86 and 2.19. The coverage of ν is 0.94.

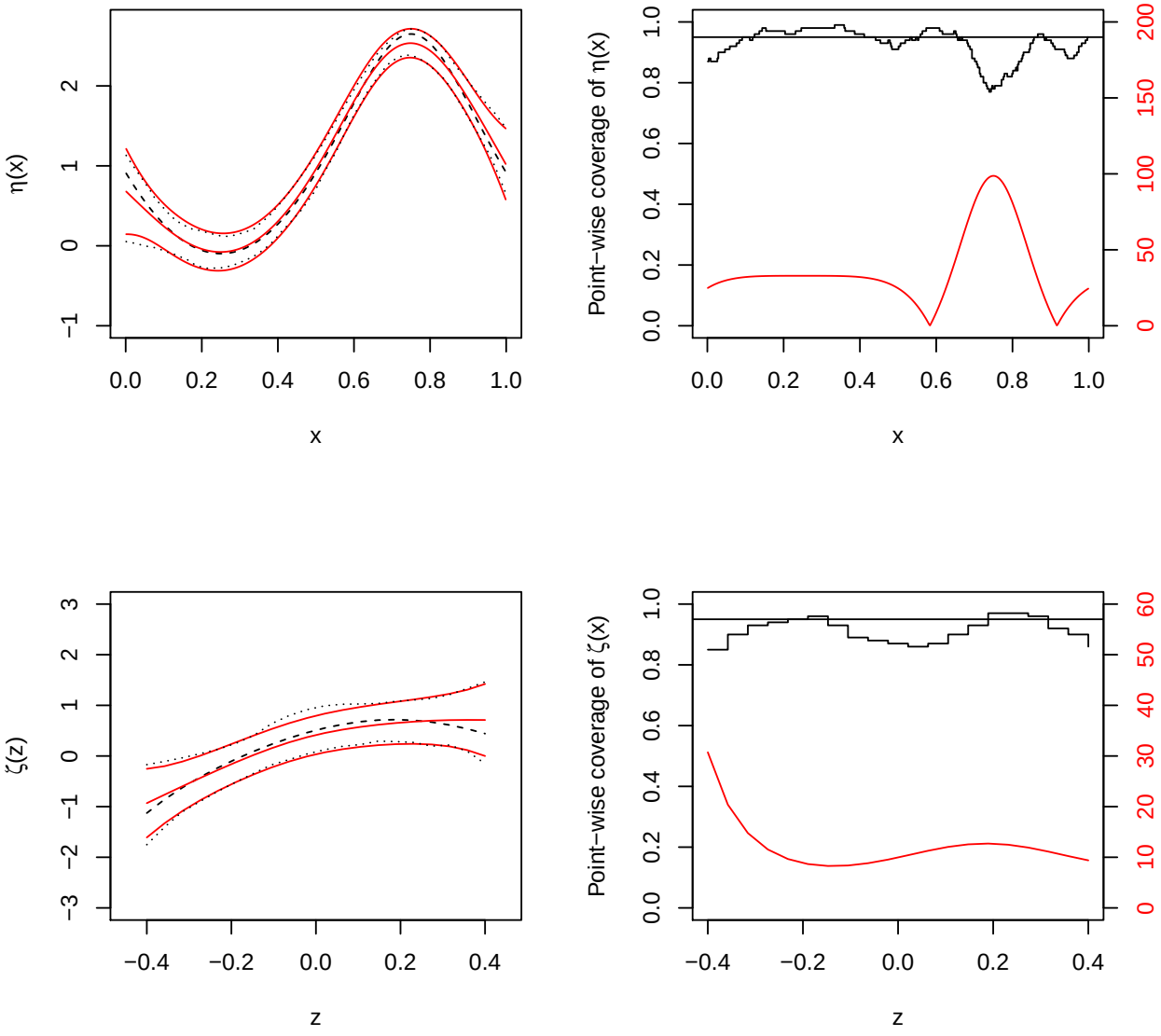

Fig. A.7. Simulation Results for Test Functions $\eta_2(\boldsymbol{x})$, $\zeta_2(\boldsymbol{z})$ and $n = 400$. Right column: Point-wise coverage (top black lines). Superimposed are nominal coverage (black straight lines) and scaled $|\eta''(x)|$ or $|\zeta''(z)|$. Left column: True functions (dashed) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (solid) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates. The average of the estimated ν is 2.06 (the true $\nu = 2$) and the average of the 95% CIs is $[1.79, 2.33]$. The empirical 2.5 and 97.5 percentiles of the estimated ν is 1.78 and 2.34. The coverage of ν is 0.94.

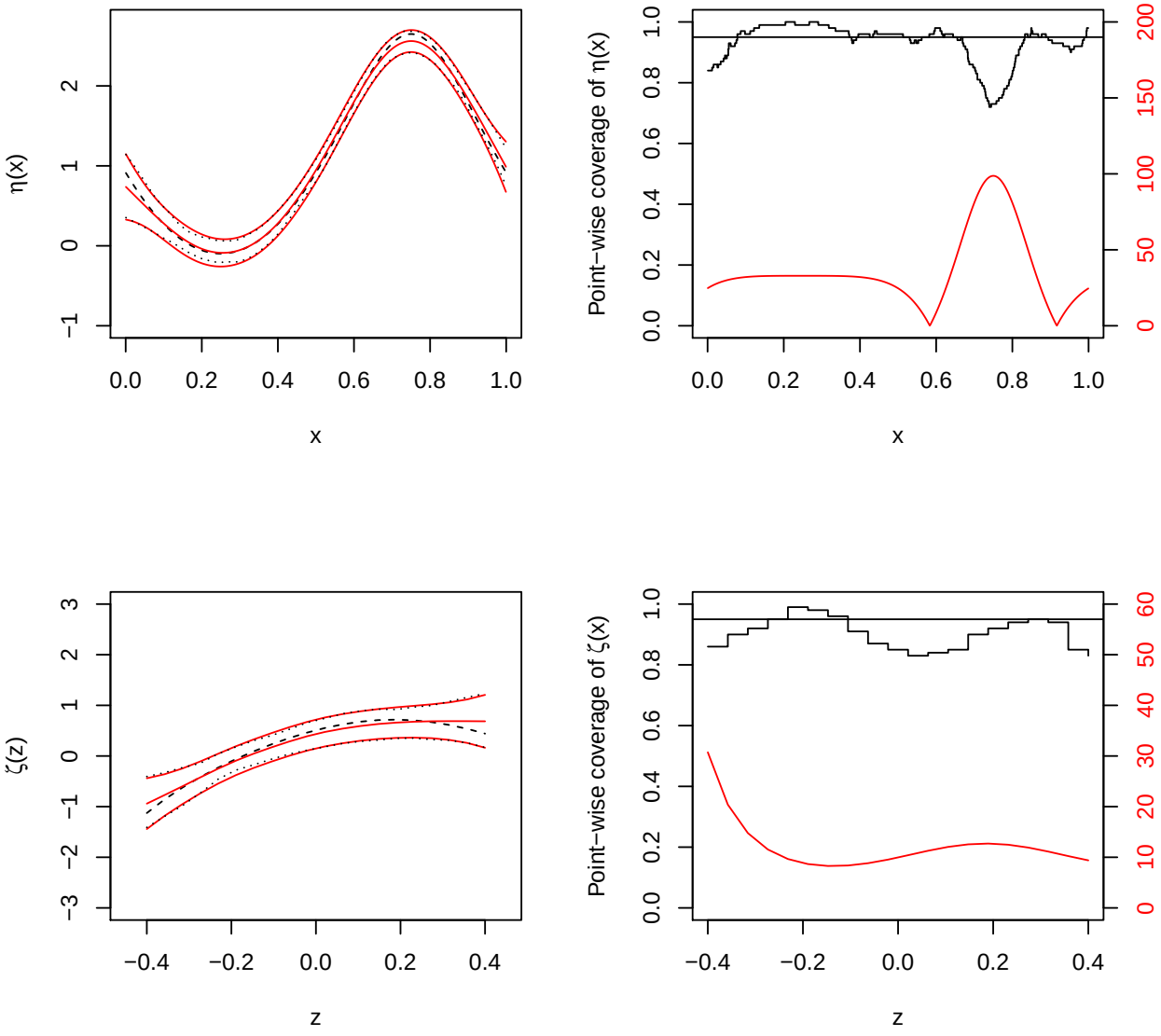

Fig. A.8. Simulation Results for Test Functions $\eta_2(\boldsymbol{x})$, $\zeta_2(\boldsymbol{z})$ and $n = 800$. Right column: Point-wise coverage (top black lines). Superimposed are nominal coverage (black straight lines) and scaled $|\eta''(x)|$ or $|\zeta''(z)|$. Left column: True functions (dashed) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (solid) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates. The average of the estimated ν is 2.02 (the true $\nu = 2$) and the average of the 95% CIs is $[1.83, 2.20]$. The empirical 2.5 and 97.5 percentiles of the estimated ν is 1.86 and 2.20. The coverage of ν is 0.95.

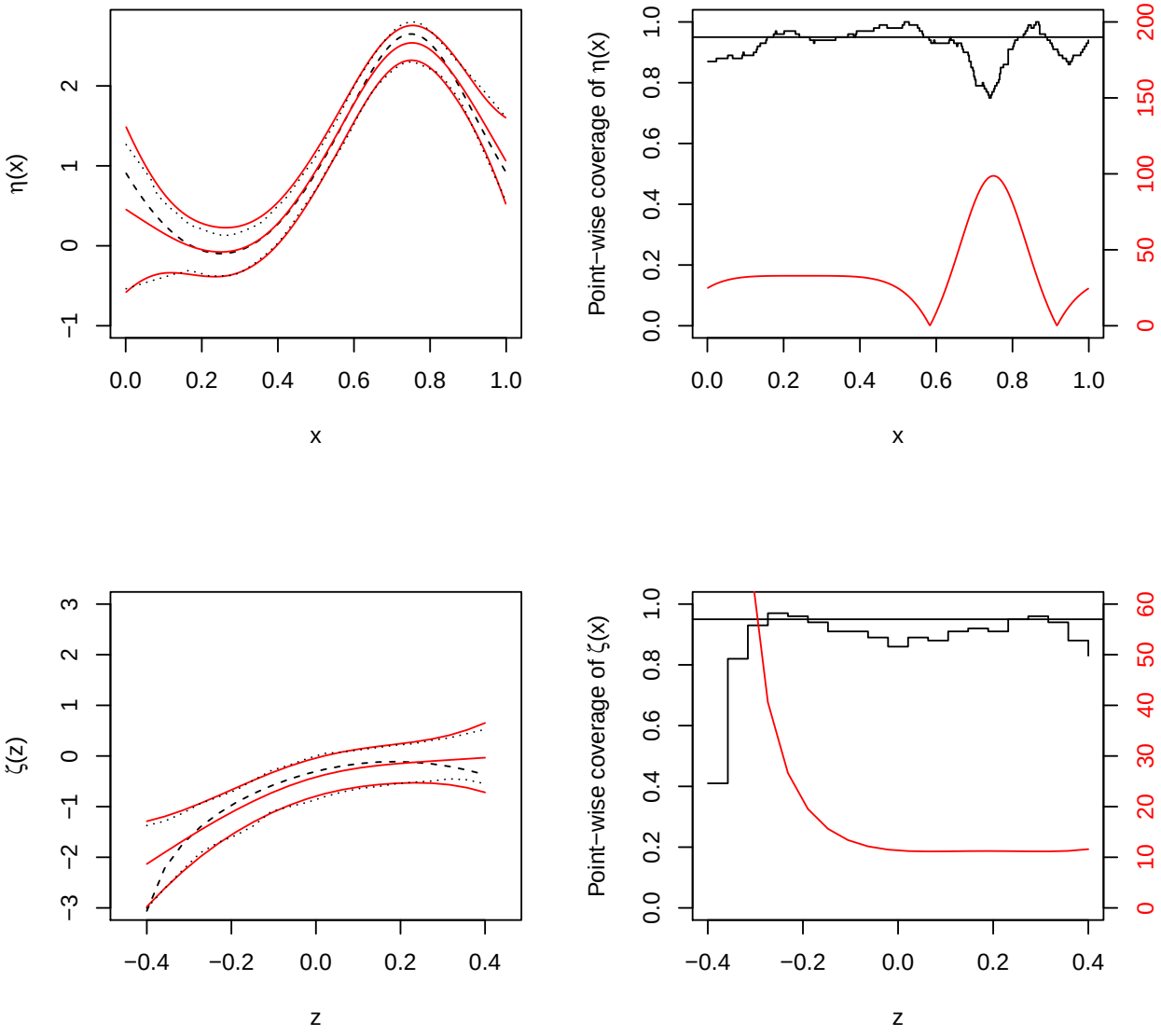

Fig. A.9. Simulation Results for Test Functions $\eta_2(\boldsymbol{x})$, $\zeta_3(\boldsymbol{z})$ and $n = 400$. Right column: Point-wise coverage (top black lines). Superimposed are nominal coverage (black straight lines) and scaled $|\eta''(x)|$ or $|\zeta''(z)|$. Left column: True functions (dashed) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (solid) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates. The average of the estimated ν is 2.07 (the true $\nu = 2$) and the average of the 95% CIs is $[1.73, 2.40]$. The empirical 2.5 and 97.5 percentiles of the estimated ν is 1.85 and 2.52. The coverage of ν is 0.95.

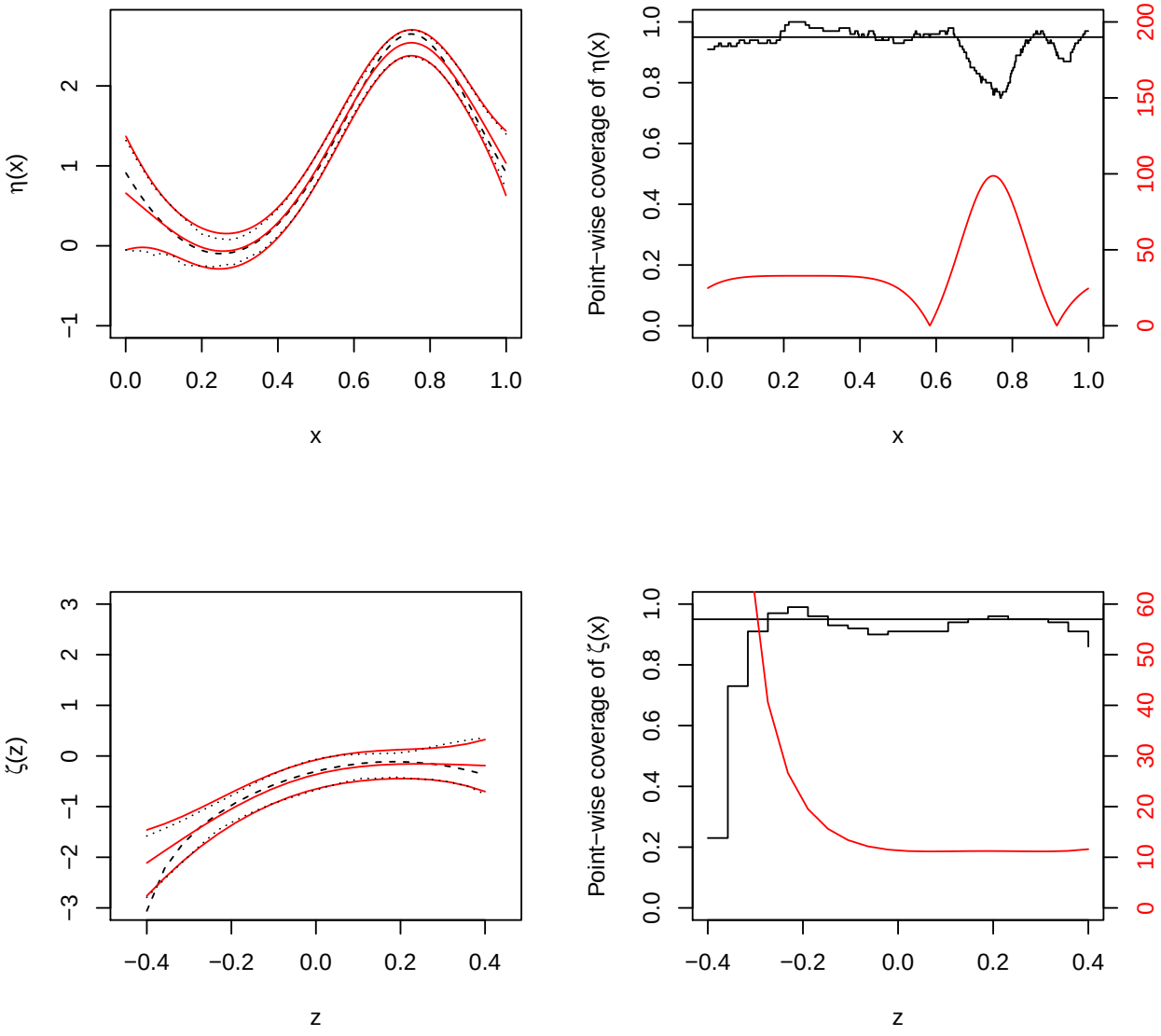

Fig. A.10. Simulation Results for Test Functions $\eta_2(\boldsymbol{x})$, $\zeta_3(\boldsymbol{z})$ and $n = 800$. Right column: Point-wise coverage (top black lines). Superimposed are nominal coverage (black straight lines) and scaled $|\eta''(x)|$ or $|\zeta''(z)|$. Left column: True functions (dashed) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (solid) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates. The average of the estimated ν is 2.04 (the true $\nu = 2$) and the average of the 95% CIs is $[1.81, 2.27]$. The empirical 2.5 and 97.5 percentiles of the estimated ν is 1.83 and 2.29. The coverage of ν is 0.93.

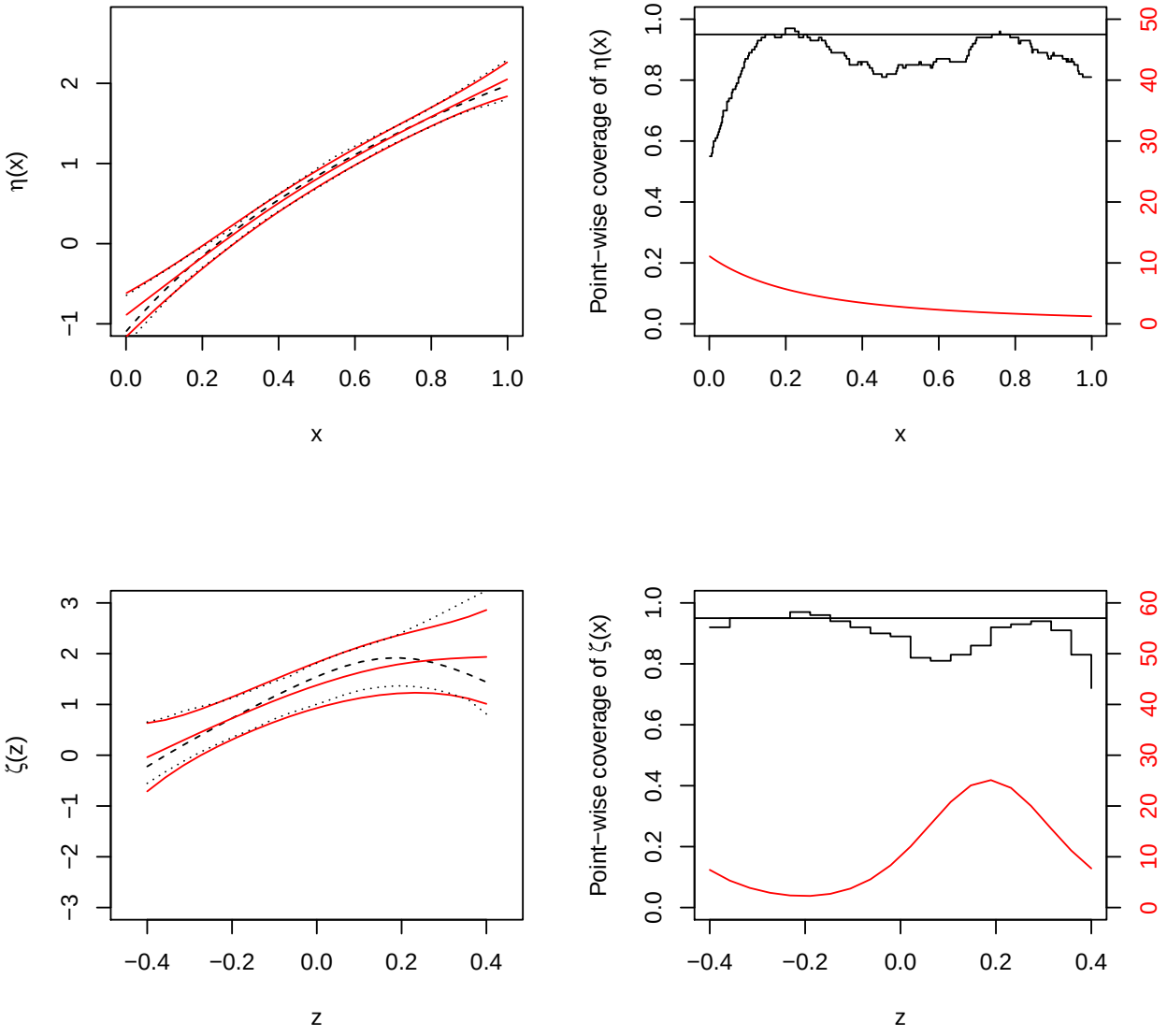

Fig. A.11. Simulation Results for Test Functions $\eta_3(\boldsymbol{x})$, $\zeta_1(\boldsymbol{z})$ and $n = 400$. Right column: Point-wise coverage (top black lines). Superimposed are nominal coverage (black straight lines) and scaled $|\eta''(x)|$ or $|\zeta''(z)|$. Left column: True functions (dashed) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (solid) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates. The average of the estimated ν is 2.03 (the true $\nu = 2$) and the average of the 95% CIs is $[1.81, 2.25]$. The empirical 2.5 and 97.5 percentiles of the estimated ν is 1.86 and 2.30. The coverage of ν is 0.95.

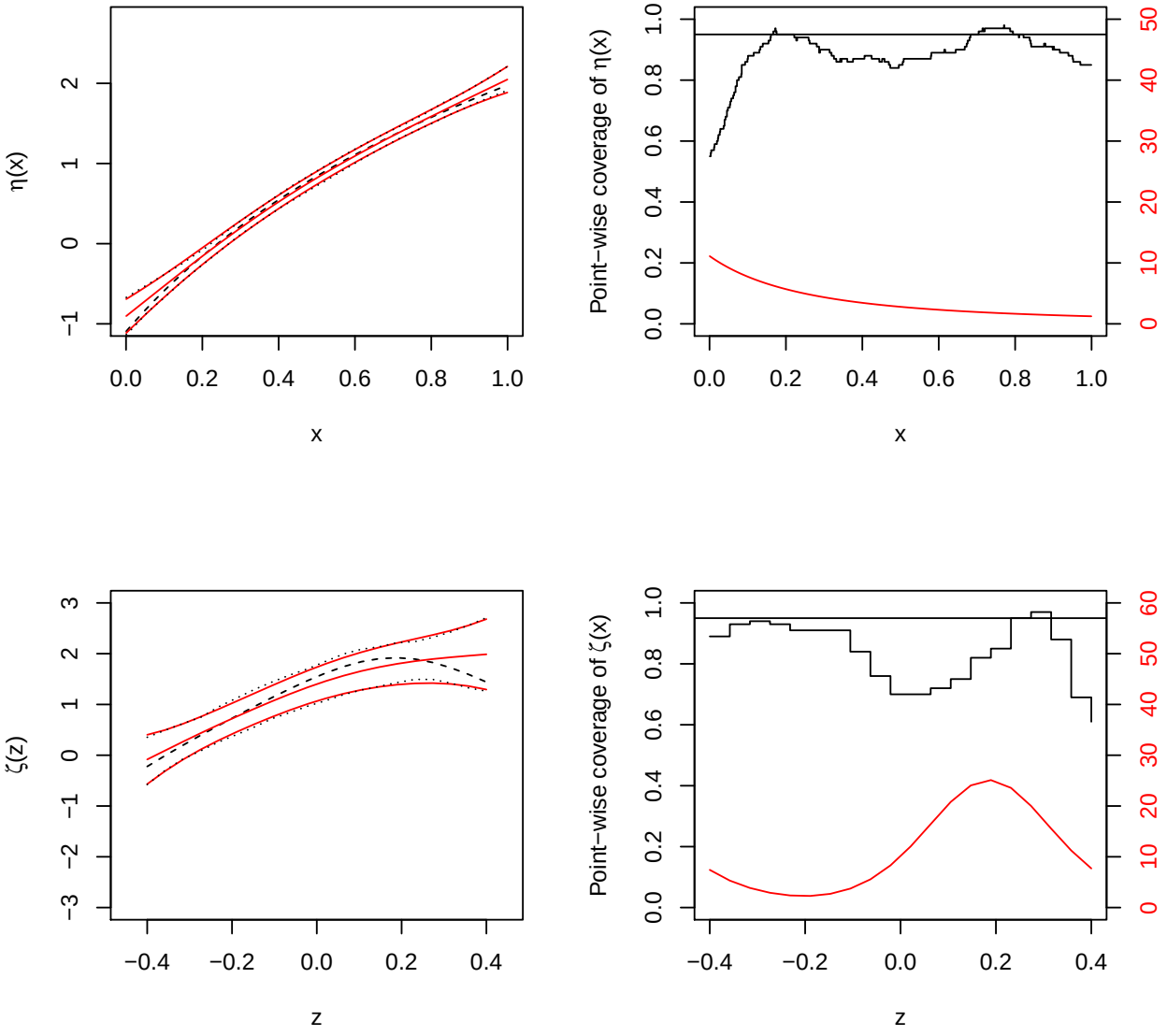

Fig. A.12. Simulation Results for Test Functions $\eta_3(\boldsymbol{x})$, $\zeta_1(\boldsymbol{z})$ and $n = 800$. Right column: Point-wise coverage (top black lines). Superimposed are nominal coverage (black straight lines) and scaled $|\eta''(x)|$ or $|\zeta''(z)|$. Left column: True functions (dashed) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (solid) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates. The average of the estimated ν is 1.99 (the true $\nu = 2$) and the average of the 95% CIs is $[1.84, 2.14]$. The empirical 2.5 and 97.5 percentiles of the estimated ν is 1.87 and 2.17. The coverage of ν is 0.95.

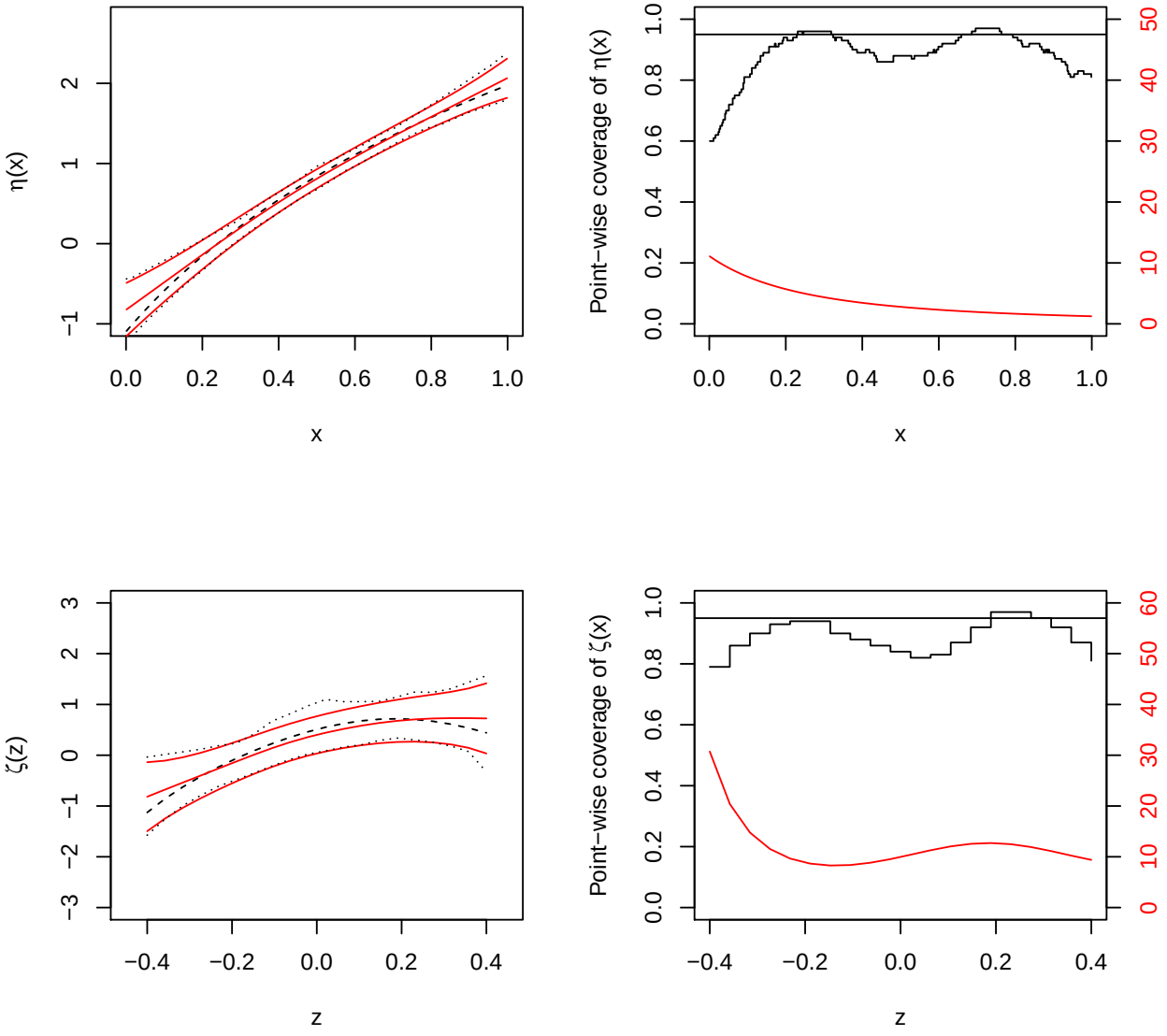

Fig. A.13. Simulation Results for Test Functions $\eta_3(\boldsymbol{x})$, $\zeta_2(\boldsymbol{z})$ and $n = 400$. Right column: Point-wise coverage (top black lines). Superimposed are nominal coverage (black straight lines) and scaled $|\eta''(x)|$ or $|\zeta''(z)|$. Left column: True functions (dashed) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (solid) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates. The average of the estimated ν is 2.04 (the true $\nu = 2$) and the average of the 95% CIs is $[1.77, 2.30]$. The empirical 2.5 and 97.5 percentiles of the estimated ν is 1.76 and 2.31. The coverage of ν is 0.94.

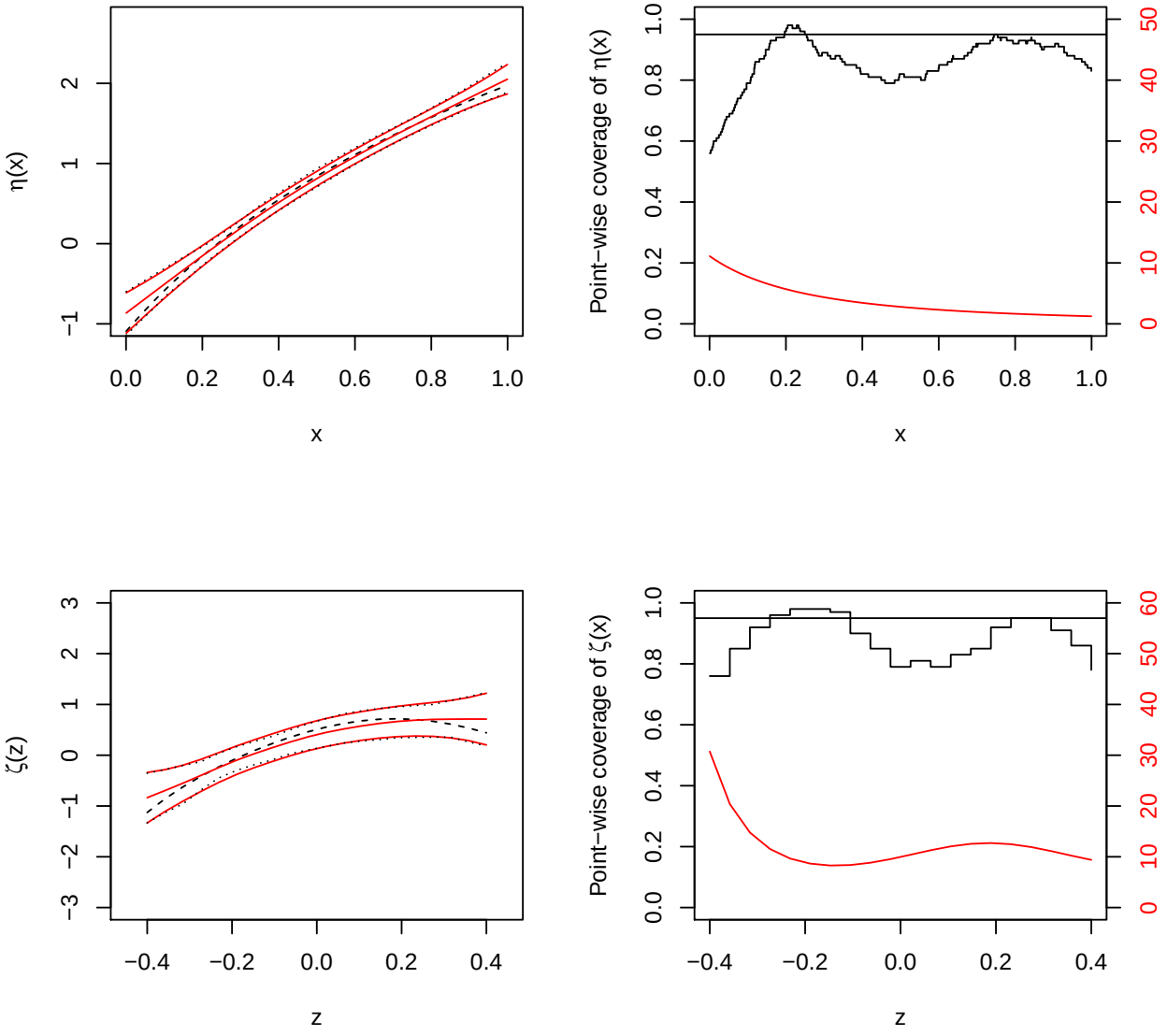

Fig. A.14. Simulation Results for Test Functions $\eta_3(\boldsymbol{x})$, $\zeta_2(\boldsymbol{z})$ and $n = 800$. Right column: Point-wise coverage (top black lines). Superimposed are nominal coverage (black straight lines) and scaled $|\eta''(x)|$ or $|\zeta''(z)|$. Left column: True functions (dashed) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (solid) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates. The average of the estimated ν is 2.00 (the true $\nu = 2$) and the average of the 95% CIs is $[1.82, 2.18]$. The empirical 2.5 and 97.5 percentiles of the estimated ν is 1.85 and 2.20. The coverage of ν is 0.95.

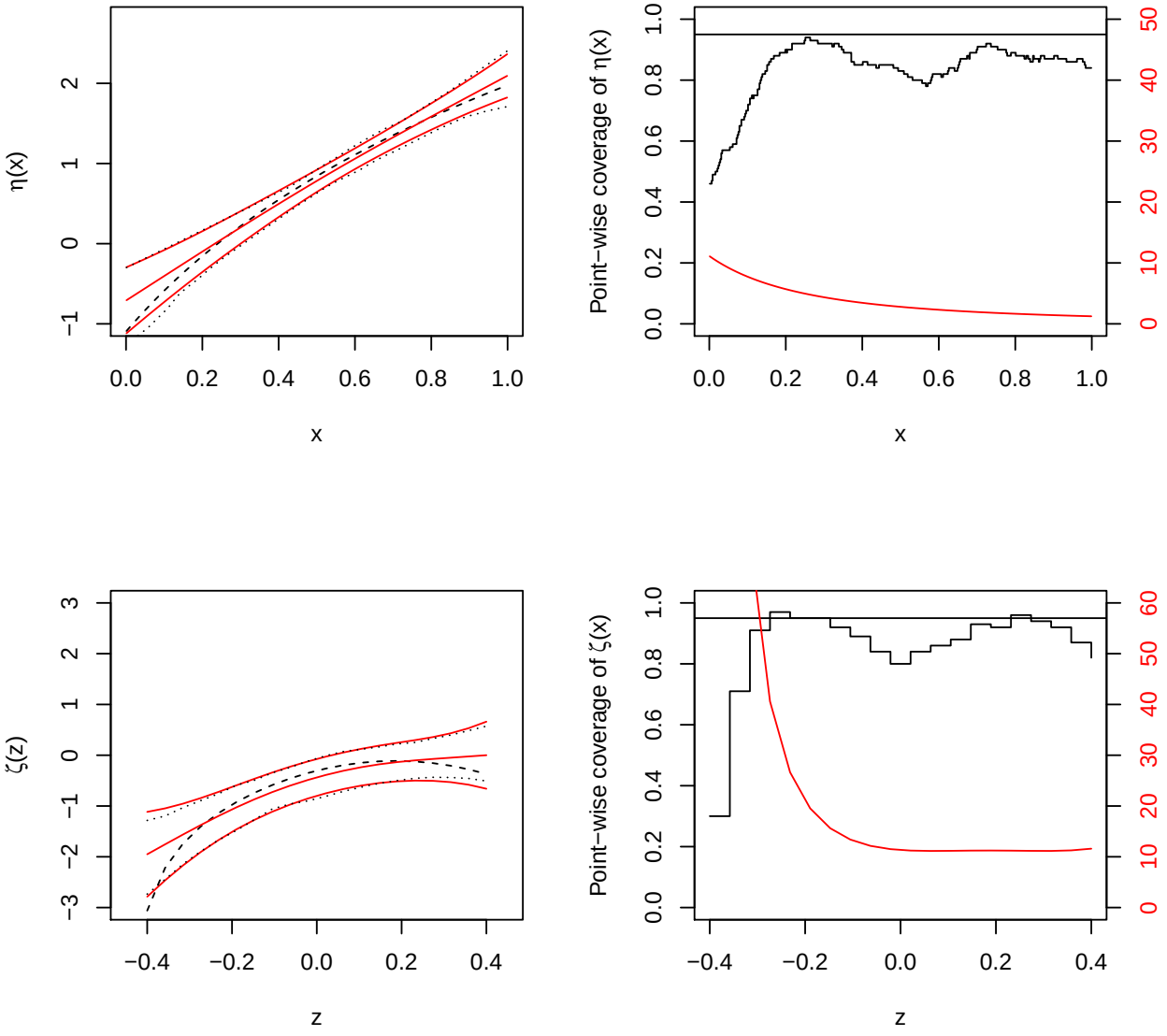

Fig. A.15. Simulation Results for Test Functions $\eta_3(\boldsymbol{x})$, $\zeta_3(\boldsymbol{z})$ and $n = 400$. Right column: Point-wise coverage (top black lines). Superimposed are nominal coverage (black straight lines) and scaled $|\eta''(x)|$ or $|\zeta''(z)|$. Left column: True functions (dashed) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (solid) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates. The average of the estimated ν is 2.03 (the true $\nu = 2$) and the average of the 95% CIs is $[1.71, 2.36]$. The empirical 2.5 and 97.5 percentiles of the estimated ν is 1.78 and 2.37. The coverage of ν is 0.97.

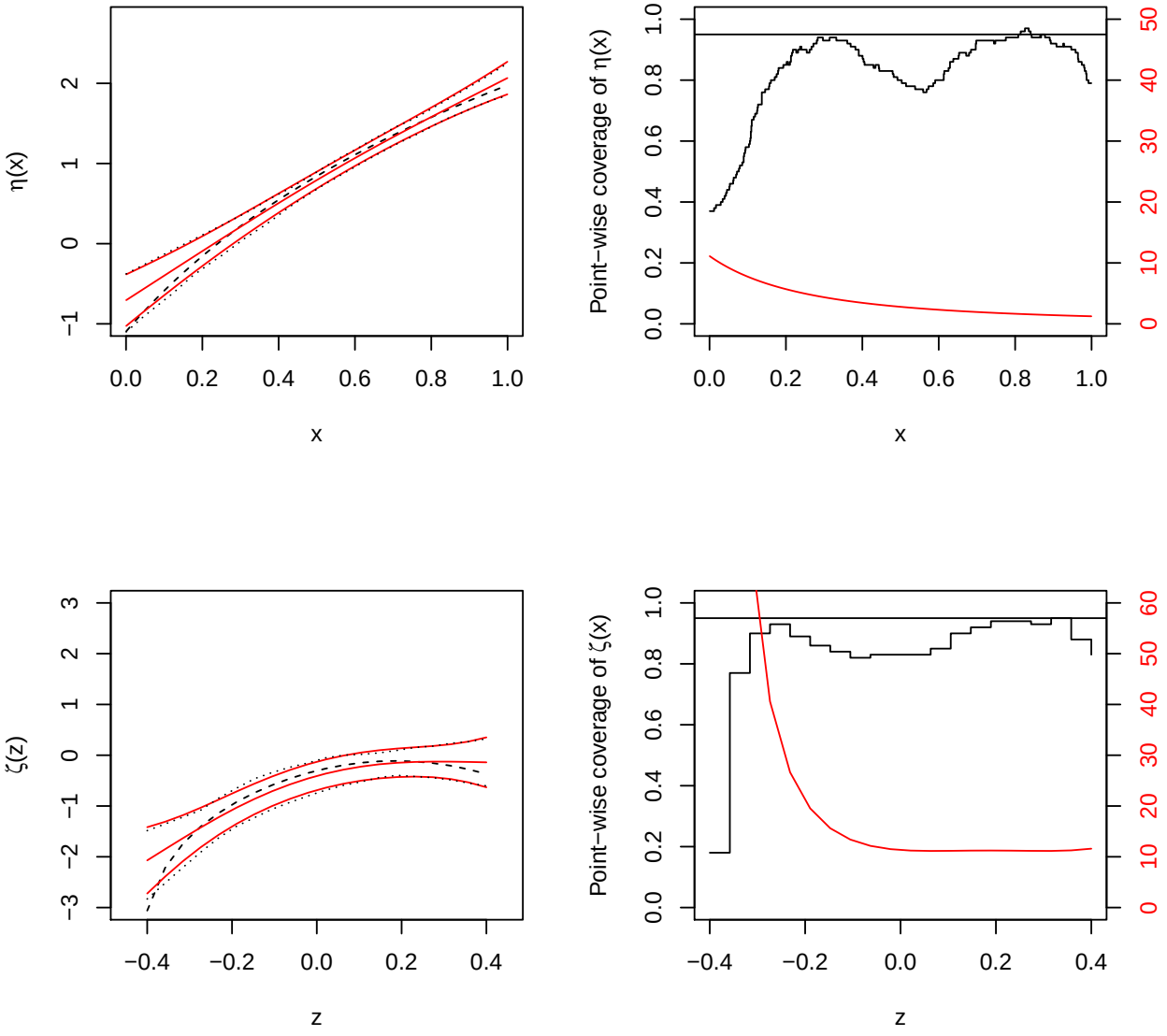

Fig. A.16. Simulation Results for Test Functions $\eta_3(\boldsymbol{x})$, $\zeta_3(\boldsymbol{z})$ and $n = 800$. Right column: Point-wise coverage (top black lines). Superimposed are nominal coverage (black straight lines) and scaled $|\eta''(x)|$ or $|\zeta''(z)|$. Left column: True functions (dashed) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (solid) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates. The average of the estimated ν is 2.00 (the true $\nu = 2$) and the average of the 95% CIs is $[1.77, 2.23]$. The empirical 2.5 and 97.5 percentiles of the estimated ν is 1.84 and 2.19. The coverage of ν is 0.99.

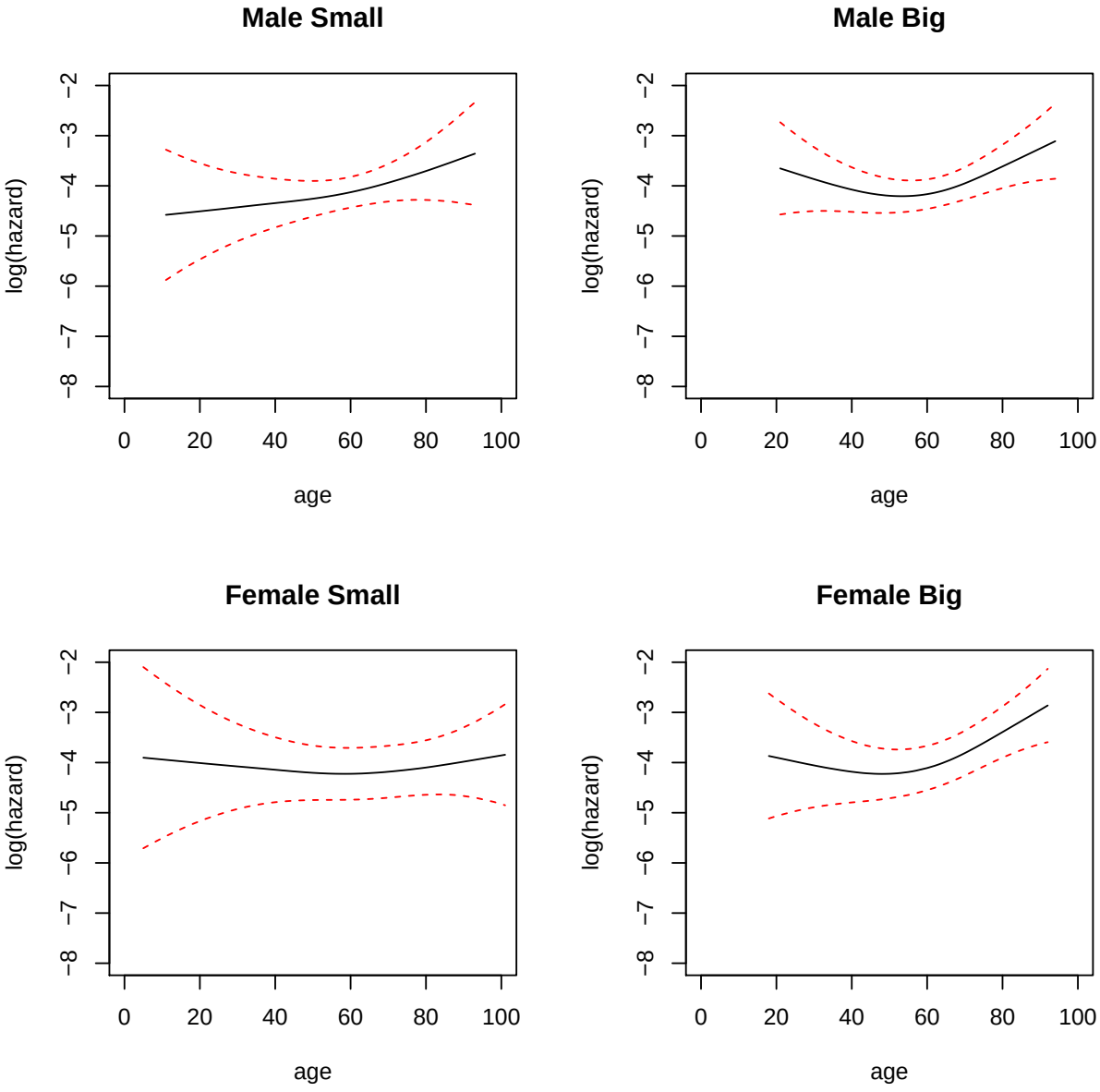

Fig. A.17. Estimated log hazard and confidence intervals against age at $time = 10$ *months*. The first row is for Male; the second row is for Female. The left column is for Small tumor size and the right column is for Big tumor size.

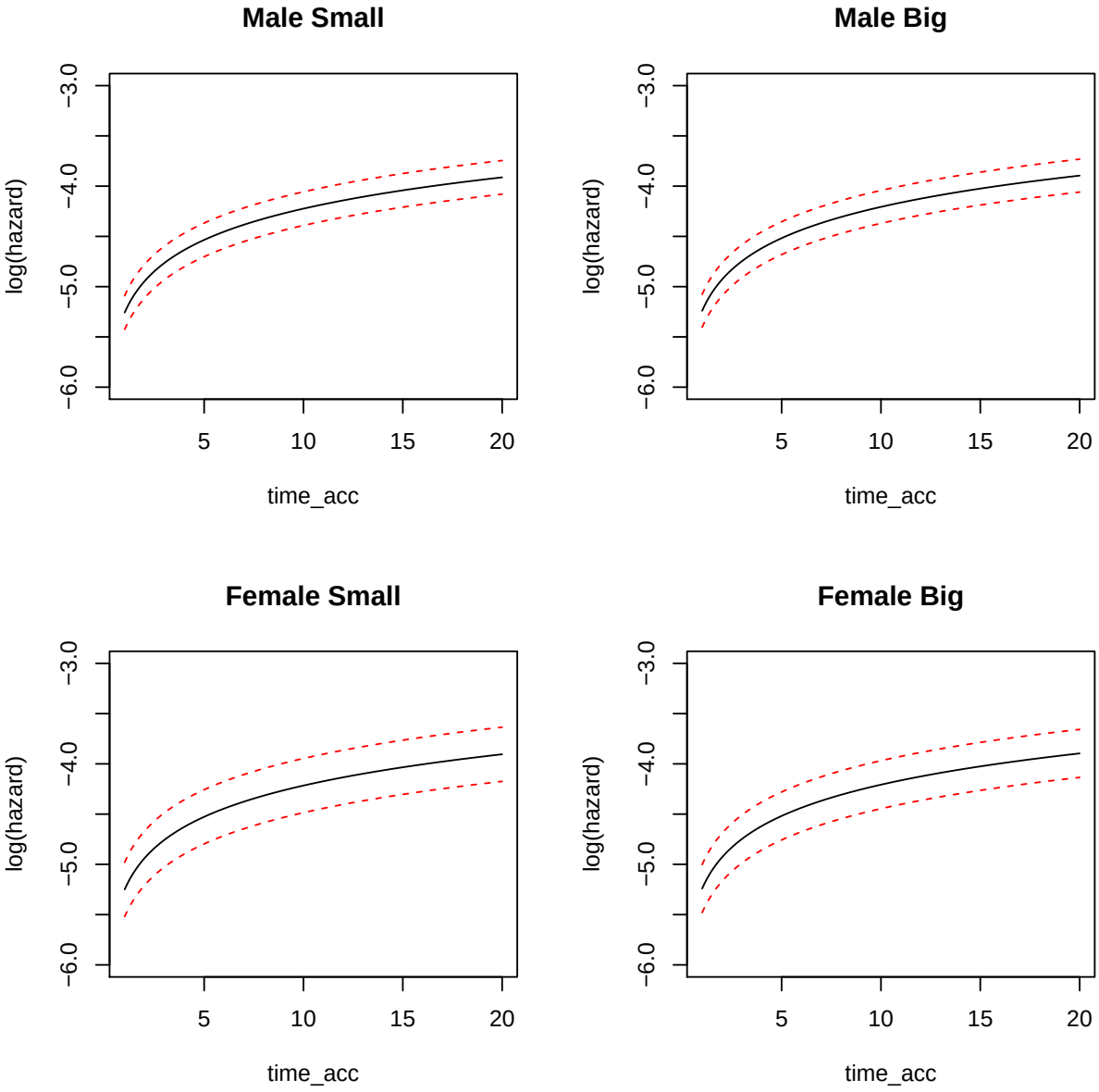

Fig. A.18. Estimated log hazard and confidence intervals against time at $age = 53\ years$. The first row is for Male; the second row is for Female. The left column is for Small tumor size and the right column is for Big tumor size.