Cure Rate Model with Spline Estimated Components

Lu Wang

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, Chairman

George R. Terrell, Co-chairman

Scotland C. Leman

Chuanhai Liu

Eric P. Smith

July 13, 2010

Blacksburg, Virginia

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

Cure Rate Model with Spline Estimated Components

Lu Wang

(ABSTRACT)

In some survival analysis of medical studies, there are often long term survivors who can be considered as permanently cured. The goals in these studies are to estimate the cure probability of the whole population and the hazard rate of the non-cured subpopulation. The existing methods for cure rate models have been limited to parametric and semiparametric models. More specifically, the hazard function part is estimated by parametric or semiparametric model where the effect of covariate takes a parametric form. And the cure rate part is often estimated by a parametric logistic regression model. We introduce a non-parametric model employing smoothing splines. It provides non-parametric smooth estimates for both hazard function and cure rate. By introducing a latent cure status variable, we implement the method using a smooth EM algorithm. Louis' formula for covariance estimation in an EM algorithm is generalized to yield point-wise confidence intervals for both functions. A simple model selection procedure based on the Kullback-Leibler geometry is derived for the proposed cure rate model. Numerical studies demonstrate excellent performance of the proposed method in estimation, inference and model selection. The application of the method is illustrated by the analysis of a melanoma study.

ACKNOWLEDGMENTS

I would like to express my immense gratitude to my major professor Professor Pang Du for his guidance, encouragement and many valuable suggestions throughout my research efforts and preparation of this dissertation. In addition to providing a very fruitful environment, his endless patience and continued interest have greatly contributed to my academic growth. I consider myself extremely fortunate to have had the opportunity to work under his direction.

I would also like to thank Professor George R. Terrell and other members of my dissertation committee, Professors Scotland Leman, Eric P. Smith and Chuanhai Liu (Purdue University). They have provided a lot of encouragement and help to me during my study at Virginia Tech.

My consulting experience has been a great resource for new ideas in my research, so I would like to thank all the faculty, staff and students who have worked with me in the consulting program, with special thanks to Professors Robert S. Schulman, Eric Vance and Ying Liu for their guidance and supervision of my consulting projects.

I feel indebted to all the faculty and staff members of the Department of Statistics for their help during my five years at Virginia Tech. Especially I want to thank Professors John P. Morgan, Inyoung Kim, Marion Reynolds and Golde I. Holtzman for their patient guidance and consistent support, the Department Head, Professor

Eric P. Smith and the Director of Graduate Programs, Professor Jeffrey B. Birch whose support and advice were very important to me. My special thanks also go to the secretaries Christina Dillon, Betty F. Higginbotham, Linda Breeding and Katie Akers, for their help, and to Mike Box, for his computer assistance, whenever I needed it.

Furthermore, I want to thank all my fellow students who have offered assistance, encouragement and friendship during the course of my study.

I must say that reaching this stage in my academic career would not have been possible without support from my parents. No words can express my deep gratitude to them.

TABLE OF CONTENTS

Page

LIST OF TABLES vii

LIST OF FIGURES viii

1 Introduction 1

1.1 Penalized Likelihood for Lifetime Data 4

1.1.1 Model Construction 5

1.1.2 Computation 8

1.2 Penalized Likelihood for Binary Data 10

1.2.1 Model Construction 11

1.2.2 Computation 12

1.3 The EM algorithm and Louis' formula 13

1.3.1 The EM algorithm 13

1.3.2 Louis' formula 18

2 Cure Rate Model with Spline Estimated Components 22

2.1 Introduction 23

2.2 Model and Identifiability 28

2.3 The EM algorithm for penalized likelihood 29

2.4 Computation of penalized likelihoods 31

2.5 Observed Information Matrix and Confidence Interval 34

2.6 Model Selection 40

2.7 Simulation Study 43

2.7.1 Estimation 43

2.7.2 Model Selection 53

2.8 Melanoma Example 59

Page

2.9 Proof 73
2.9.1 Proof of Proposition 2.2.1 (i) 73
2.9.2 Proof of Proposition 2.2.1 (ii) 74
2.9.3 Proof of 2.21 76
LIST OF REFERENCES 78
A Simulation 83

LIST OF TABLES

Table Page

2.1 Model Selection Simulation Results. 0.05 is used as the cutoff value for the ρ statistics in the simulation study. 55

2.2 Model selection of the cure rate component of melanoma example. ρ is the model selection statistics. A ρ value smaller than 0.05 indicates the reduced model. 67

2.3 Model selection of the survival component of melanoma example. ρ is the model selection statistics. A ρ value smaller than 0.05 indicates the reduced model. 68

LIST OF FIGURES

Figure Page

2.1 The four test hazard surfaces. 48

2.2 Simulation Results for Test Functions $\pi_2(z)$, $h_1(t,x)$ and $n = 800$. Left column: Point-wise coverages (top black lines). Superimposed are nominal coverage (red lines) and scaled $|logit(\pi''(z))|$ (bottom black line) or histogram of data. Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates. 49

2.3 Simulation Results for Test Functions $\pi_3(z)$, $h_1(t,x)$ and $n = 800$. Left column: Point-wise coverages (top black lines). Superimposed are nominal coverage (red lines) and scaled $|logit(\pi''(z))|$ (bottom black line) or histogram of data. Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates. 50

2.4 Simulation Results for Test Functions $\pi_2(z)$, $h_1(t,x)$ and $n = 400$. Left column: Point-wise coverages (top black lines). Superimposed are nominal coverage (red lines) and scaled $|logit(\pi''(z))|$ (bottom black line) or histogram of data. Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates. 51

2.5 Simulation Results for Test Functions $\pi_3(z)$, $h_1(t,x)$ and $n = 400$. Left column: Point-wise coverages (top black lines). Superimposed are nominal coverage (red lines) and scaled $|logit(\pi''(z))|$ (bottom black line) or histogram of data. Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates. 52

2.6 Model selection simulation setting 1. Circle: models (t,x) vs. $(t,x,x_c,x*x_c)$. Square: models (t) vs. (t,x) 56

Figure Page

2.7 Model selection simulation setting 2. Circle: models (t, x_c) vs. $(t, x, x_c, x * x_c)$. Square: models (t) vs. (t, x) 57

2.8 Model selection simulation setting 3. Square: models (z_c) vs. (z, z_c). Diamond: models (z) vs. (z, z_c). Circle: models (z, z_c) vs. $(z, z_c, z * z_c)$. 58

2.9 Plot of data. Black circles are observed failures, red circles are observed censoring. 62

2.10 Estimated logit cure rates and their confidence intervals against age. Size: S=small B=big. Gender: M=male F=female. Superimposed are true data points with positions determined by age and converged $y's$. Black circles are observed failures, red circles are observed censoring. 63

2.11 Estimated log hazard and confidence intervals against time at $age = 53\ years$. Size: S=small B=big. Gender: M=male F=female. 64

2.12 Estimated log hazard and confidence intervals against age at $time = 10\ months$. Size: S=small B=big. Gender: M=male F=female. 65

2.13 Estimated logit cure rates and their confidence intervals against age. Superimposed are true data points with positions determined by age and converged $y's$. Black circles are observed failures, red circles are observed censoring. 69

2.14 Estimated log hazard surfaces. 70

2.15 Estimated log hazard and confidence intervals against time at $age = 53\ years$. 71

2.16 Estimated log hazard and confidence intervals against age at $time = 10\ months$. 72

A.1 Simulation Results for Test Functions $\pi_4(z)$, $h_5(t, x)$ and $n = 400$. Left column: Point-wise coverages (top black lines). Superimposed are nominal coverage (red lines) and scaled $|logit(\pi''(z))|$ (bottom black line) or histogram of data. Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates. 84

Figure Page

A.2 Simulation Results for Test Functions $\pi_1(z)$, $h_1(t,x)$ and $n = 400$. Left column: Point-wise coverages (top black lines). Superimposed are nominal coverage (red lines) and scaled $|logit(\pi''(z))|$ (bottom black line) or histogram of data. Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates. 85

A.3 Simulation Results for Test Functions $\pi_1(z)$, $h_2(t,x)$ and $n = 400$. Left column: Point-wise coverages (top black lines). Superimposed are nominal coverage (red lines) and scaled $|logit(\pi''(z))|$ (bottom black line) or histogram of data. Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates. 86

A.4 Simulation Results for Test Functions $\pi_2(z)$, $h_2(t,x)$ and $n = 400$. Left column: Point-wise coverages (top black lines). Superimposed are nominal coverage (red lines) and scaled $|logit(\pi''(z))|$ (bottom black line) or histogram of data. Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates. 87

A.5 Simulation Results for Test Functions $\pi_3(z)$, $h_2(t,x)$ and $n = 400$. Left column: Point-wise coverages (top black lines). Superimposed are nominal coverage (red lines) and scaled $|logit(\pi''(z))|$ (bottom black line) or histogram of data. Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates. 88

A.6 Simulation Results for Test Functions $\pi_1(z)$, $h_3(t,x)$ and $n = 400$. Left column: Point-wise coverages (top black lines). Superimposed are nominal coverage (red lines) and scaled $|logit(\pi''(z))|$ (bottom black line) or histogram of data. Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates. 89

Figure Page

A.7 Simulation Results for Test Functions $\pi_2(z)$, $h_3(t,x)$ and $n = 400$. Left column: Point-wise coverages (top black lines). Superimposed are nominal coverage (red lines) and scaled $|logit(\pi''(z))|$ (bottom black line) or histogram of data. Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates. 90

A.8 Simulation Results for Test Functions $\pi_3(z)$, $h_3(t,x)$ and $n = 400$. Left column: Point-wise coverages (top black lines). Superimposed are nominal coverage (red lines) and scaled $|logit(\pi''(z))|$ (bottom black line) or histogram of data. Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates. 91

A.9 Simulation Results for Test Functions $\pi_1(z)$, $h_4(t,x)$ and $n = 400$. Left column: Point-wise coverages (top black lines). Superimposed are nominal coverage (red lines) and scaled $|logit(\pi''(z))|$ (bottom black line) or histogram of data. Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates. 92

A.10 Simulation Results for Test Functions $\pi_2(z)$, $h_4(t,x)$ and $n = 400$. Left column: Point-wise coverages (top black lines). Superimposed are nominal coverage (red lines) and scaled $|logit(\pi''(z))|$ (bottom black line) or histogram of data. Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates. 93

A.11 Simulation Results for Test Functions $\pi_3(z)$, $h_4(t,x)$ and $n = 400$. Left column: Point-wise coverages (top black lines). Superimposed are nominal coverage (red lines) and scaled $|logit(\pi''(z))|$ (bottom black line) or histogram of data. Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates. 94

Figure Page

A.12 Simulation Results for Test Functions $\pi_1(z)$, $h_1(t,x)$ and $n = 800$. Left column: Point-wise coverages (top black lines). Superimposed are nominal coverage (red lines) and scaled $|logit(\pi''(z))|$ (bottom black line) or histogram of data. Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates. 95

A.13 Simulation Results for Test Functions $\pi_1(z)$, $h_2(t,x)$ and $n = 800$. Left column: Point-wise coverages (top black lines). Superimposed are nominal coverage (red lines) and scaled $|logit(\pi''(z))|$ (bottom black line) or histogram of data. Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates. 96

A.14 Simulation Results for Test Functions $\pi_2(z)$, $h_2(t,x)$ and $n = 800$. Left column: Point-wise coverages (top black lines). Superimposed are nominal coverage (red lines) and scaled $|logit(\pi''(z))|$ (bottom black line) or histogram of data. Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates. 97

A.15 Simulation Results for Test Functions $\pi_3(z)$, $h_2(t,x)$ and $n = 800$. Left column: Point-wise coverages (top black lines). Superimposed are nominal coverage (red lines) and scaled $|logit(\pi''(z))|$ (bottom black line) or histogram of data. Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates. 98

A.16 Simulation Results for Test Functions $\pi_1(z)$, $h_3(t,x)$ and $n = 800$. Left column: Point-wise coverages (top black lines). Superimposed are nominal coverage (red lines) and scaled $|logit(\pi''(z))|$ (bottom black line) or histogram of data. Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates. 99

Figure Page

A.17 Simulation Results for Test Functions $\pi_2(z)$, $h_3(t,x)$ and $n = 800$. Left column: Point-wise coverages (top black lines). Superimposed are nominal coverage (red lines) and scaled $|logit(\pi''(z))|$ (bottom black line) or histogram of data. Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates. 100

A.18 Simulation Results for Test Functions $\pi_3(z)$, $h_3(t,x)$ and $n = 800$. Left column: Point-wise coverages (top black lines). Superimposed are nominal coverage (red lines) and scaled $|logit(\pi''(z))|$ (bottom black line) or histogram of data. Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates. 101

A.19 Simulation Results for Test Functions $\pi_1(z)$, $h_4(t,x)$ and $n = 800$. Left column: Point-wise coverages (top black lines). Superimposed are nominal coverage (red lines) and scaled $|logit(\pi''(z))|$ (bottom black line) or histogram of data. Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates. 102

A.20 Simulation Results for Test Functions $\pi_2(z)$, $h_4(t,x)$ and $n = 800$. Left column: Point-wise coverages (top black lines). Superimposed are nominal coverage (red lines) and scaled $|logit(\pi''(z))|$ (bottom black line) or histogram of data. Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates. 103

A.21 Simulation Results for Test Functions $\pi_3(z)$, $h_4(t,x)$ and $n = 800$. Left column: Point-wise coverages (top black lines). Superimposed are nominal coverage (red lines) and scaled $|logit(\pi''(z))|$ (bottom black line) or histogram of data. Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates. 104

1. INTRODUCTION

In some survival studies the population under consideration consists of two groups of subjects, susceptible and non-susceptible subjects. The susceptible subjects are not cured and would eventually experience failures, while non-susceptible subjects are not at risk of developing such events and can be regarded as cured. Data from such studies are often referred as cure rate data. One popular category of cure rate model is promotion cure model. It is popular for Bayesian cure survival analysis and has an attractive biological interpretation, see, e.g., Yakovlev and Tsodikov [1996], Tsodikov [1998], Chen et al. [1999], Zeng et al. [2006]. Although useful in some cure rate studies, the special requirement in the form of the model prevents a possible extension to nonparametric estimation. Thus, we don't pursue this direction here. Berkson and Gage [1952] first proposed two-component mixture cure model to analyze cure rate data. The model was developed assuming the study population was a mixture of two components: a survival component and a cure rate component. 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. Recently, 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. 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 both cure rate and hazard functions 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 settings relevant to our cure rate data problem are hazard estimation and logistic regression, so in this chapter we introduce the smoothing spline estimation details for these two settings, see Wahba [1990] and Gu [2004].

The function estimates in our cure rate problem 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 Lifetime Data

Censored lifetime data are common in life testing, medical follow-up and other studies. 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, 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 $S_0(t, u) = \text{Prob}(T > t|U = u)$. Of interest is the estimation of the hazard function $h_0(t, u) = -\partial \log S_0(t, u)/\partial t$ or its logarithm $\eta_0(t, u) = \log h_0(t, u)$.

The contribution of subject i to the likelihood is $h_0(X_i, U_i)^{\delta_i} S_0(X_i, U_i)/S_0(Z_i, U_i)$, that is $S_0(C_i, U_i)/S_0(Z_i, U_i)$ when the subject is censored, and $h_0(T_i, U_i)S_0(T_i, U_i)/S_0(Z_i, U_i)$ otherwise. Hence the penalized likelihood in this problem estimates η_0 by minimizing

$$-\frac{1}{n}\sum_{i=1}^{n}\{\delta_i \eta(X_i, U_i) - \int_{Z_i}^{X_i} e^{\eta(t,U_i)} dt\} + \frac{\lambda}{2} J(\eta) \tag{1.2}$$

in a Hilbert space $\mathcal{H}$ of functions on the product domain of time and covariate. This formulation, found in **?** along with an asymptotic theory, evolved from the work of Anderson and Senthilselvan [1980], O'Sullivan [1988a], O'Sullivan [1988b], and Zucker and Karr [1990], among others. The minimizer of (1.2) is actually the maximum likelihood estimate (MLE) under a soft constraint of the form $J(\eta) \le \rho$ for some $\rho > 0$, with λ being the Lagrange multiplier; see, e.g., Section 2.6.2 in Gu [2002]. With $\lambda = \infty$ (i.e. $\rho = 0$), one enforces a parametric model in the null space of

$J(\eta)$, $\{\eta : J(\eta) = 0\}$, and as $\lambda \to 0$ (i.e. $\rho \to \infty$), one approaches the nonparametric MLE; the latter is the Kaplan-Meier in the absence of the covariate U. The practical performance of the estimate hinges on the proper selection of λ, for which an effective cross-validation procedure can be found in Section 7.2 of Gu [2002].

1.1.1 Model Construction

In this section, we describe how the penalty term $J(\eta)$ and the covariate in (1.2) specify a model. The minimization of (1.2) is done in a Hilbert space $\mathcal{H}$ of functions on the product domain $\mathcal{T} \times \mathcal{U}$ of time and covariate. $J(\eta)$ is taken as a square seminorm in $\mathcal{H}$ with a finite dimensional null space $\mathcal{N}_J \subset \mathcal{H}$, where a finite dimensional $\mathcal{N}_J$ prevents interpolation, the conceptual equivalence of a delta sum. The evaluation functional $[t, u]f = f(t, u)$ is assumed to be continuous in $f \in \mathcal{H}$, which is necessary for (1.2) to be continuous in its argument η. When $\mathcal{U}$ is a singleton (i.e., with no covariate), the formulation reduces to that of O'Sullivan [1988a].

A space $\mathcal{H}$ in which the evaluation functional is continuous is called a reproducing kernel Hilbert space (RKHS) possessing a reproducing kernel (RK) $R(\cdot, \cdot)$, a non-negative definite function satisfying $R_x(\cdot) = R(x, \cdot) \in \mathcal{H}$, $\forall x = (t, u) \in \mathcal{T} \times \mathcal{U}$, and $\langle R(x, \cdot), f(\cdot)\rangle = f(x)$, $\forall f \in \mathcal{H}$, where $\langle \cdot, \cdot \rangle$ is the inner product in $\mathcal{H}$; the RK $R(\cdot, \cdot)$ and the space $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ determine each other uniquely. Typically, $\langle \cdot, \cdot \rangle = J(\cdot, \cdot) + \tilde{J}(\cdot, \cdot)$, where $J(\cdot, \cdot)$ is the semi inner product associated with $J(\cdot)$ and $\tilde{J}(\cdot, \cdot)$ is an inner product in the null space $\mathcal{N}_J = \{\eta : J(\eta) = 0\}$ of $J(\eta)$ when restricted therein. There exists a tensor sum decomposition $\mathcal{H} = \mathcal{N}_J \oplus \mathcal{H}_J$, where the space

$\mathcal{H}_J$ has $J(\eta)$ as its square norm and an RK R_J satisfying $J(R_J(x,\cdot), f(\cdot)) = f(x)$, $\forall f \in \mathcal{H}_J$. See, e.g., Section 2.1 in Gu [2002].

Next, we give some examples of RKHS configurations under different covariate settings. Without loss of generality, let us assume the time domain $\mathcal{T} = [0,1]$ in all the examples in this section.

Example 1.1.1 (Singleton $\mathcal{U}$) *A singleton $\mathcal{U}$ indicates the absence of a covariate. One only has the time domain $\mathcal{T} = [0,1]$. A choice of $J(\eta)$ is $\int_0^1 (\eta'')^2 dt$, which yields the popular cubic splines. A choice of $\tilde{J}(f,g)$ is $(\int_0^1 f\,dt)(\int_0^1 g\,dt) + (\int_0^1 \dot{f}\,dt)(\int_0^1 \dot{g}\,dt)$, yielding $\mathcal{H}_J = \{\eta : \int_0^1 \eta dt = \int_0^1 \dot{\eta} dt = 0, J(\eta) < \infty\}$ and the RK $R_J(t_1,t_2) = k_2(t_1)k_2(t_2) - k_4(t_1 - t_2)$, where $k_\nu = B_\nu/\nu!$ are scaled Bernoulli polynomials. The null space $\mathcal{N}_J$ has a basis $\{1, k_1(t)\}$, where $k_1(t) = t - 0.5$. See, e.g., Section 2.3.3 in Gu [2002].* □

Example 1.1.2 (Doubleton $\mathcal{U}$) *A doubleton $\mathcal{U}$, say $\mathcal{U} = \{1,2\}$, gives 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* $\mathcal{H}_{\langle t\rangle}$ *on the time domain*

$$\begin{aligned}\mathcal{H}_{\langle t\rangle} &= \left\{\eta : \int_0^1 (\eta'')^2 dx < \infty\right\} = \mathcal{H}_{00\langle t\rangle} \oplus \mathcal{H}_{01\langle t\rangle} \oplus \mathcal{H}_{1\langle t\rangle} \\ &= span\{1\} \oplus span\{k_1(x)\} \oplus \left\{\eta : \int_0^1 \eta dx = \int_0^1 \dot{\eta} dx = 0,\ \int_0^1 (\eta'')^2 dx < \infty\right\},\end{aligned}$$

with RKs $R_{00\langle t\rangle}(t_1,t_2)=1$, $R_{01\langle t\rangle}(t_1,t_2)=k_1(t_1)k_1(t_2)$, *and* $R_{1\langle t\rangle}(t_1,t_2)=k_2(t_1)k_2(t_2)-k_4(t_1-t_2)$. *Taking tensor product of* $\mathcal{H}_{\langle t\rangle}$ *and* $\mathcal{H}_{\langle u\rangle}$, *one obtains six tensor sum terms* $\mathcal{H}_{\nu,\mu}=\mathcal{H}_{\nu\langle t\rangle}\otimes\mathcal{H}_{\mu\langle u\rangle}$ *on* $\mathcal{T}\times\mathcal{U}$, $\nu=00,01,1$ *and* $\mu=0,1$, *with RKs* $R_{\nu,\mu}(x_1,x_2)=R_\nu(t_1,t_2)R_\mu(u_1,u_2)$, *where* $x_i=(t_i,u_i)$. *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*

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

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

$$\eta(t,u)=\eta_\emptyset+\eta_t(t)+\eta_u(u)+\eta_{t,u}(t,u)$$

for functions on $\mathcal{T} \times \mathcal{U}$*, with* $\eta_\emptyset \in \mathcal{H}_{00\langle t\rangle} \otimes \mathcal{H}_{0\langle u\rangle}$ *being the constant term,* $\eta_t \in \{\mathcal{H}_{01\langle t\rangle} \oplus \mathcal{H}_{1\langle t\rangle}\} \otimes \mathcal{H}_{0\langle u\rangle}$ *the* t *main effect,* $\eta_u \in \mathcal{H}_{00\langle t\rangle} \otimes \mathcal{H}_{1\langle u\rangle}$ *the* u *main effect, and* $\eta_{t,u} \in \{\mathcal{H}_{01\langle t\rangle} \oplus \mathcal{H}_{1\langle t\rangle}\} \otimes \mathcal{H}_{1\langle u\rangle}$ *the interaction. One may obtain an additive model for log hazard* $\eta(t, u)$*, i.e., a proportional hazard model, by setting* $\eta_{t,u} = 0$*. See, e.g., Example 2.8 in Gu [2002].*

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 (Continuous Interval $\mathcal{U}$) *A continuous interval* $\mathcal{U}$*, say* $\mathcal{U} = [0, 1]$*, describes a univariate continuous covariate. Now the RKHS* $\mathcal{H}_{\langle u\rangle}$ *on the* u *domain can have the same cubic spline decomposition as that of* $\mathcal{H}_{\langle t\rangle}$ *in Example 1.1.2, and its tensor product with* $\mathcal{H}_{\langle t\rangle}$ *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

Let $N = \sum_{i=1}^{n} \delta_i$ be the number of events and (X_i^*, U_i^*), $i = 1, \ldots, N$, be the observed lifetimes along with the associated covariates. The space $\mathcal{H}$ is usually infinite dimensional, and the minimizer of (1.2) in $\mathcal{H}$ is in general not computable. To

circumvent the problem, **?** proposed to use the minimizer of (1.2) in an adaptive finite dimensional space $\mathcal{H}_N = \mathcal{N}_J \oplus \text{span}\{R_J((X_i^*, U_i^*), \cdot), i = 1, \ldots, N\}$. Under mild conditions, the minimizer of (1.2) in $\mathcal{H}_N$ was shown to share the same asymptotic convergence rates as the minimizer in $\mathcal{H}$.

Write $\xi_j = R_J((X_j^*, U_j^*), \cdot)$ and let $\{\phi_\nu\}_{\nu=1}^m$ be a basis of $\mathcal{N}_J$. By definition, a function in $\mathcal{H}_N$ has an expression

$$\eta = \sum_{\nu=1}^{m} d_\nu \phi_\nu + \sum_{j=1}^{N} c_j \xi_j = \boldsymbol{\phi}^T \mathbf{d} + \boldsymbol{\xi}^T \mathbf{c}, \tag{1.3}$$

where $\boldsymbol{\phi}$ and $\boldsymbol{\xi}$ are vectors of functions and $\mathbf{d}$ and $\mathbf{c}$ are vectors of coefficients. Substituting (1.3) into (1.2), noting that

$$J(\eta) = \big\langle \sum_{j=1}^{N} c_j \xi_j \, , \, \sum_{k=1}^{N} c_k \xi_k \big\rangle = \sum_{j=1}^{N} \sum_{k=1}^{N} c_j c_k R_J((X_j^*, U_j^*), (X_k^*, U_k^*)),$$

one calculates the minimizer η_λ of (1.2) in $\mathcal{H}_N$ by minimizing

$$-\frac{1}{n} \mathbf{1}^T (S\mathbf{d} + R\mathbf{c}) + \frac{1}{n} \sum_{i=1}^{n} \int_{Z_i}^{X_i} \exp(\boldsymbol{\phi}_i^T \mathbf{d} + \boldsymbol{\xi}_i^T \mathbf{c}) dt + \frac{\lambda}{2} \mathbf{c}^T Q \mathbf{c} \tag{1.4}$$

with respect to $\mathbf{d}$ and $\mathbf{c}$, where S is $N \times m$ with the (i, ν)th entry $\phi_\nu(X_i^*, U_i^*)$, R is $N \times N$ with the (i, j)th entry $\xi_j(X_i^*, U_i^*) = R_J((X_i^*, U_i^*), (X_j^*, U_j^*))$, Q is $N \times N$ with the (j, k)th entry $\xi_j(X_k^*, U_k^*) = R_J((X_j^*, U_j^*), (X_k^*, U_k^*))$, $\boldsymbol{\phi}_i$ is $m \times 1$ with the νth entry $\phi_\nu(t, U_i)$, and $\boldsymbol{\xi}_i$ is $N \times 1$ with the jth entry $\xi_j(t, U_i)$.

Write $\mu_f(g) = n^{-1}\sum_{i=1}^{n}\int_{Z_i}^{X_i} g(t,U_i)e^{f(t,U_i)}dt$ and $V_f(g,h) = \mu_f(gh)$. The minimization of (1.4) for fixed smoothing parameters can be done through Newton iteration, which updates the coefficients from the current iterate $\tilde{\eta} = \boldsymbol{\phi}^T\tilde{\mathbf{d}} + \boldsymbol{\xi}^T\tilde{\mathbf{c}}$ through

$$\begin{pmatrix} V_{\phi,\phi} & V_{\phi,\xi} \\ V_{\xi,\phi} & V_{\xi,\xi} + \lambda Q \end{pmatrix} \begin{pmatrix} \mathbf{d} \\ \mathbf{c} \end{pmatrix} = \begin{pmatrix} S^T\mathbf{1}/n - \mu_\phi + V_{\phi,\eta} \\ R^T\mathbf{1}/n - \mu_\xi + V_{\xi,\eta} \end{pmatrix}, \tag{1.5}$$

where $V_{\phi,\phi} = V_{\tilde{\eta}}(\boldsymbol{\phi},\boldsymbol{\phi}^T)$, $V_{\phi,\xi} = V_{\xi,\phi}^T = V_{\tilde{\eta}}(\boldsymbol{\phi},\boldsymbol{\xi}^T)$, $V_{\xi,\xi} = V_{\tilde{\eta}}(\boldsymbol{\xi},\boldsymbol{\xi}^T)$, $\mu_\phi = \mu_{\tilde{\eta}}(\boldsymbol{\phi})$, $\mu_\xi = \mu_{\tilde{\eta}}(\boldsymbol{\xi})$, $V_{\phi,\eta} = V_{\tilde{\eta}}(\boldsymbol{\phi},\tilde{\eta})$, and $V_{\xi,\eta} = V_{\tilde{\eta}}(\boldsymbol{\xi},\tilde{\eta})$; see, e.g., Section 7.1 of Gu [2002]. The selection of the smoothing parameters can be done through an outer-loop optimization of a cross-validation score derived in Section 7.2 of Gu [2002].

1.2 Penalized Likelihood for Binary Data

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

$$f(y|x) = exp\{(y\eta(x) - b(\eta(x)))/a(\phi) + c(y,\phi)\} \tag{1.6}$$

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 (1.6) includes Gaussian, Binomial, Poisson distributions as special cases. The penalized likelihood for estimating η is

$$-\frac{1}{n}\sum_{i=1}^{n}\{Y_i\eta(x_i) - b(\eta(x_i))\} + \frac{\lambda}{2}J(\eta). \tag{1.7}$$

Note that in (1.7), 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

$$\sum_{i=1}^{n} \left\{ y_i \eta(x_i) - \log[1 + e^{\eta(x_i)}] \right\}, \tag{1.8}$$

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

1.2.1 Model Construction

Note that η in (1.8) is a restriction free function like the log hazard eta in Section 1.1. Hence, as in Section 1.1, we don't need any other constraints on eta besides the smoothness requirement imposed by the penalty $J(\cdot)$. Consequently, the construction of the RKHS and the corresponding model terms for binary data is similar to that for hazard estimation in Section 1.1.1. Next, we give brief descriptions of RKHS configurations under different covariate settings for binary data.

Example 1.2.1 (One covariate $\mathcal{V}$) *Equating the one covariate $\mathcal{V}$ with $\mathcal{T}$ in Example 1.1.1, the construction of RKHS is the same as Example 1.1.1.* □

Example 1.2.2 (One continuous and one categorical covariates $\mathcal{V} \times \mathcal{U}$) *Equating the one continuous covariate $\mathcal{V}$ and the one categorical covariate $\mathcal{U}$ with $\mathcal{T}$ and $\mathcal{U}$ in*

Example 1.1.2 respectively, the construction of RKHS is the same as Example 1.1.2. □

Example 1.2.3 (Two continuous covariates $\mathcal{V} \times \mathcal{U}$) *Equating the two continuous covariates $\mathcal{V}$ and $\mathcal{U}$ with $\mathcal{T}$ and $\mathcal{U}$ in Example 1.1.3 respectively, the construction of RKHS is the same as Example 1.1.3.* □

Example 1.2.4 (Multi-dimensional $\mathcal{U}$) *A multi-dimensional $\mathcal{U}$ corresponds to a multivariate covariate U. Clearly the tensor product structures in Examples 1.2.2 and 1.2.3 can be augmented to accommodate the additional dimensions.* □

1.2.2 Computation

The first term of (1.7) 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 (1.7) has an expression

$$\eta = \sum_{\nu=1}^{m} d_\nu \phi_\nu + \sum_{j=1}^{N} c_j \xi_j = \boldsymbol{\phi}^T \mathbf{d} + \boldsymbol{\xi}^T \mathbf{c}, \tag{1.9}$$

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 (1.7) 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

$$-\frac{1}{2}\tilde{w}_i \{\eta(x_i) - \tilde{\eta}(x_i) + \tilde{u}_i/\tilde{w}_i\}^2 + C_i \tag{1.10}$$

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

$$-\frac{1}{n}\sum_{i=1}^{n} \tilde{w}_i(\tilde{Y}_i - \eta(x_i))^2 + \lambda J(\eta) \tag{1.11}$$

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

1.3 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:

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

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

$$f(x, y|\theta) = b(x, y)exp\{\theta t(x, y)^T\}/a(\theta) \tag{1.12}$$

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 re-

construction, 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 conceptual 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:

$$S^*(x,\theta) = E_\theta[S(x,Y,\theta)],$$

$$S^*(x,\widehat{\theta}) = 0,$$

$$I_{obs}(\theta) = E_\theta[B(x,Y,\theta)] - E_\theta[S(x,Y,\theta)S(x,Y,\theta)^T] + S^*(x,\theta)S^{*T}(x,\theta) \qquad (1.13)$$

The first term in (1.13) 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:*

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

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. CURE RATE MODEL WITH SPLINE ESTIMATED COMPONENTS

This Chapter proposes a nonparametric estimation procedure for cure rate data based on penalized likelihood method. In some survival analysis of medical studies, there are often long term survivors who can be considered as permanently cured. The goals in these studies are to estimate the cure probability of the whole population and the hazard rate of the non-cured subpopulation. When covariate is present as often happens in practice, the understanding of covariate effects on the cure probability and hazard rate is of equal importance. The existing methods are limited to parametric and semiparametric models. We propose a two-component mixture cure rate model with nonparametric forms for both the cure probability and the hazard rate function. Identifiability of the model is guaranteed by an additive assumption on hazard rate. Estimation is carried out by an EM algorithm for maximizing a penalized likelihood. For inferential purposes, we extend the Louis formula to obtain point-wise confidence intervals for cure probability and hazard rate. Model selection procedure is also developed to identify negligible components in a functional ANOVA decomposition of the proposed cure rate model. We then evaluate the proposed method by extensive simulations. An application to a melanoma study demonstrates the method.

2.1 Introduction

Lifetime data are common in clinical trials, industrial life testing and other studies. In some survival studies the population under consideration consists of two groups of subjects, susceptible and non-susceptible individuals. All susceptible subjects would eventually experience the failure if there is no censoring, while non-susceptible subjects are not at risk of developing such events and can be regarded as cured. Examples include cancer studies with long-term survivors, smoking studies with permanent quitters, employment studies with secured employees. Among them, cancer studies with long-term survivors are of particular interest for researchers in public health. Modern cancer treatments have substantially improved cure rates and have generated a great interest in and need for proper statistical tools to analyze survival data with nonnegligible cure fractions. Research on many cancer types, including prostate, breast, colon and ovarian cancer, non-Hodgkins lymphoma, leukemia, and melanoma have shown that a significant proportion of patients with these cancers are cured after therapy Tai et al. [2005]; that is, an individual will have little or no risk of experiencing the event of interest (e.g., breast cancer) after treatment. The focus of cure rate model is on the estimation of the proportion of non-susceptible individuals and the failure time distribution of the susceptible individuals at the same time.

One popular category of cure rate model is promotion cure model. It is popular for Bayesian cure survival analysis and has an attractive biological interpretation, see, e.g., Yakovlev and Tsodikov [1996], Tsodikov [1998], Chen et al. [1999],

Zeng et al. [2006]. In their model, the survival function for the whole population is $S(t,x) = G[-\theta(x,\beta)F(t)]$, 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. 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 the exponential function 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 $\exp(-\theta(x,\beta))$. Although useful in some cure rate studies, the special requirement in the form of the model prevents a possible extension to nonparametric estimation. Thus, we don't pursue this direction here.

The model we consider belongs to another popular category of models for cure rate data, called two-component mixture cure model. It is developed assuming the study population is a mixture of two components and has a survival function

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

where $\pi(\boldsymbol{z})$ and $S(t|\boldsymbol{x})$ are respectively the proportion and the survival function of susceptible subjects; $\boldsymbol{z}$ and $\boldsymbol{x}$ are the associated covariates. The promotion cure model has one covariate function that describes both the survival and cure components. In contrast, the two components of the mixture cure model can depend on different

parameters, allowing for separate covariate interpretations for the cure function and for the survival function of those who are not cured.

Two-component mixture cure model was first proposed in Berkson and Gage [1952]. In that study, $\pi(\boldsymbol{z})$ was an unknown constant without incorporating covariates. The $S(t|\boldsymbol{x})$ was modeled parametrically. In Farewell [1982], $\pi(\boldsymbol{z})$ was modeled as parametric logistic regression model and $S(t|\boldsymbol{x})$ assumed Weibull distribution. Instead of assuming Weibull distribution, Kuk and Chen [1992] extended Farewell [1982] by formulating $S(t|\boldsymbol{x})$ 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. The difference between Peng and Dear [2000] and Sy and Taylor [2000] is that Peng and Dear [2000] still based on a marginal likelihood approach but Sy and Taylor [2000] applied a full likelihood approach. Recently, 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. Included as special cases are the proportional hazards cure model and the proportional odds cure model, the latter of which specifies the survival distribution from the proportional odds regression model. There, $\pi(\boldsymbol{z})$ was modeled as parametric logistic regression and $S(t|\boldsymbol{x})$ was modeled semiparametrically. In Othus et al. [2009], the semiparametric transformation model that allows for covariates as well as dependent censoring was proposed. The key idea is to use an inverse censoring probability reweighting scheme to derive unbiased esti-

mating equations that account for dependent censoring. In this manner, they are able to avoid making parametric assumptions about the dependence structure between the survival time and the censoring time. It also is noteworthy that their proposed model, which accommodates time-dependent covariates, is more general than that proposed by Lu and Ying [2004], which allows for only time-independent covariates.

The drawback of the existing approaches for cure rate model is that they are limited to parametric and semiparametric models. More specifically, the hazard function of susceptible subjects takes either a parametric form such as Weibull distribution or a semiparametric form such as the Cox proportional hazards model where the log relative risk is linear in covariate. And the cure rate part is in essence a parametric logistic regression model where a pre-specified link function of cure rate is linear in covariates. In practice, such parametric or semiparametric assumptions may not hold and the analysis tools thus derived may not be valid. For example, some clinical studies involve covariates that have "optimal dose", indicating a nonlinear pattern, which is difficult to be captured by these parametric and semiparametric models.

To address the limitation of the existing parametric and semiparametric models, we propose in this paper a more flexible nonparametric model based on penalized likelihood employing smoothing spline. It provides non-parametric smooth estimates for both cure rate and survival components. The cure rate $\pi(\boldsymbol{z})$ is modeled as a non-parametric logistic regression. The survival component $S(t|\boldsymbol{x})$ is modeled as a non-parametric proportional hazard model in terms of smooth hazard rate $\eta(t|\boldsymbol{x})$ where both baseline hazard and covariates are modeled non-parametrically. By intro-

ducing a latent cure status variable, we implement the method using a smooth EM algorithm. Louis' formula for covariance estimation in an EM algorithm is generalized to yield point-wise confidence intervals for both functions. A simple model selection procedure based on the Kullback-Leibler geometry is derived for the proposed cure rate model. Numerical studies demonstrate excellent performance of the proposed method in estimation, inference and model selection. The application of the method is illustrated by the analysis of a melanoma study.

As far as we know, this is the first purely nonparametric method ever proposed for cure rate data. In additional to its flexibility in modeling, our method offers smooth function estimates that are appealing to practitioners especially at the exploratory stage of data analysis. The confidence intervals developed here also provide reliable inference tools necessary in public health studies. The model selection tool allows reliable specifications of the model terms in the functional ANOVA decomposition associated with smoothing splines.

The rest of the work is organized as follows. In Section 2.2, we give a detailed description of our nonparametric cure rate model and then address the identifiability of the model in our nonparametric setting. In Section 2.3, we propose an EM algorithm to estimate the unknown functions. In Section 2.4, computation of penalized likelihoods is addressed. The observed information matrix and confidence interval are derived in Section 2.5. In Sections 2.6, model selection tool is developed. In Sections 2.7 and 2.8, we present simulation studies and an application to a real example respectively.

2.2 Model and Identifiability

We first introduce some notation. For each subject i, one observes $(t_i, \delta_i, \boldsymbol{z}_i, \boldsymbol{x}_i)$, $i = 1, \ldots, n$. Here $t_i = \min(T_i, C_i)$ with T_i being the survival time and C_i being the right censoring time, $\delta_i = I_{T_i \leq C_i}$ the indicator of failure, $\boldsymbol{z}_i$ is the covariate associated with cure rate $\pi(\cdot)$ and $\boldsymbol{x}_i$ is the covariate associated with hazard function $h(\cdot, \cdot)$. Note that all the cured subjects are censored and have $\delta_i = 0$, but some censored subjects may experience failures beyond the study period.

We consider the *two-component mixture cure model* (2.1) for cure rate data. Assuming non-informative censoring, the observed likelihood function following the two-component mixture cure model formulation can be written as

$$l_{\text{obs}}(\pi(\cdot), S(\cdot)) \propto \prod_{i=1}^{n} [\pi(\boldsymbol{z}_i) f(t_i, \boldsymbol{x}_i)]^{\delta_i} [1 - \pi(\boldsymbol{z}_i) + \pi(\boldsymbol{z}_i) S(t, \boldsymbol{x}_i)]^{1-\delta_i}, \tag{2.2}$$

where $f(t, \boldsymbol{x})$ and $S(t, \boldsymbol{x})$ are respectively the probability density function and the survival function of failure time t given covariate $\boldsymbol{x}$. This likelihood is very difficult for direct optimization because the cure rate and survival components are entangled. In Section 2.3, we will introduce a latent curing status variable y and use an EM algorithm to optimize the penalized complete likelihood.

Model identifiability can be an issue in statistical models for cure rate data. Extending Li et al. [2001], we prove a proposition on identifiability conditions. One sufficient condition for the model to be identifiable is that the hazard function for

susceptible subjects has a proportional hazard structure or equivalently, the log hazard function has an additive structure.

Proposition 2.2.1 *(i) The model defined by*

$$S_{pop}(t, z, x) = 1 - \pi(z) + \pi(z)[S(t, x)]$$

where $\pi(z)$ and $S_0(t, x)$ are unspecified, is not identifiable.

(ii) The model defined by

$$S_{pop}(t, z, x) = 1 - \pi(z) + \pi(z)[S(t)]^{r(x)}$$

where $\pi(z)$, $r(x)$ and $S(t)$ are unspecified, is identifiable if $r(x) \in R$, $x \in [0, 1]^p$.

2.3 The EM algorithm for penalized likelihood

Direct optimization of the likelihood (2.2) is difficult. A popular approach to address this problem is to introduce a latent curing status variable y and to apply the EM algorithm; see, e.g., Peng and Dear [2000] Sy and Taylor [2000]. We now extend the classical EM algorithm to the optimization of penalized likelihood. Let y be the unobservable indicator of non-curing such that $y_i = 1$ if the ith subject is not cured and $y_i = 0$ if cured. Given $\boldsymbol{y} = (y_1, \ldots, y_n)$, the complete log likelihood, in terms of

$\zeta(\boldsymbol{z}) = \log[\pi(\boldsymbol{z})/(1-\pi(\boldsymbol{z}))]$ and the log hazard function $\eta(t,x) = \log(f(t,x)/S(t,x))$, can be written as $L_c(\zeta, \eta; \boldsymbol{y}) = L_1(\zeta; \boldsymbol{y}) + L_2(\eta; \boldsymbol{y})$, where

$$L_1(\zeta; \boldsymbol{y}) = \sum_{i=1}^{n} \big\{ y_i \zeta(\boldsymbol{z}_i) - \log[1 + e^{\zeta(\boldsymbol{z}_i)}] \big\}, \tag{2.3}$$

$$L_2(\eta; \boldsymbol{y}) = \sum_{i=1}^{n} \big\{ y_i \delta_i \eta(t_i, \boldsymbol{x}_i) - y_i \int_0^{t_i} e^{\eta(t, \boldsymbol{x}_i)} dt \big\}. \tag{2.4}$$

Notice the complete log likelihood decomposes into a sum of two separated parts with each part involving only one component. Thus, one can optimize with respect to ζ and η separately with y acting as the bridge between the two parts.

In the E-step, one computes the conditional expectation of L_c with respect to the latent variable y_i's given the current estimates $\Theta^{(m)} = \{\zeta^{(m)}, \eta^{(m)}\}$. Let

$$y_i^{(m)} = E[y_i | \Theta^{(m)}] = \delta_i + (1 - \delta_i) \frac{\exp[-\int_0^{t_i} e^{\eta(t, \boldsymbol{x}_i)} dt]}{\exp[-\zeta(\boldsymbol{z}_i)] + \exp[-\int_0^{t_i} e^{\eta(t, \boldsymbol{x}_i)} dt]} \Bigg|_{\Theta^{(m)}}.$$

The M-step then minimizes

$$-\frac{1}{n} L_1(\zeta; \boldsymbol{y}^{(m)}) + \frac{\beta}{2} J(\zeta) \tag{2.5}$$

and

$$-\frac{1}{n} L_2(\eta; \boldsymbol{y}^{(m)}) + \frac{\lambda}{2} J(\eta) \tag{2.6}$$

to obtain $\Theta^{(m+1)} = \{\zeta^{(m+1)}, \eta^{(m+1)}\}$.

Computation is often a major concern when the EM algorithm is used. In our simulation study, the EM algorithm often converges in less than 10 steps. This faster convergence with a little smoothing at the M-step was demonstrated in [43], where they proposed a smoothed EM (EMS) algorithm for mixture density estimation. Note that both of the objective functionals in the M-step are concave. Their optimizations can be easily handled by, e.g., the Newton-Raphson procedure. For the convergence of the whole EM algorithm, the inventors of the algorithm suggested in their seminal paper [14] that maximizing a penalized likelihood, hence obtaining a smoother estimate, in the M-step would actually yield a quicker convergence than maximizing the original likelihood. At the end of each EM iteration of EMS, it adds an extra smoothing step to smooth the new estimate before feeding it to the next E-step. The EM algorithm proposed in this project follows the same heuristic and converges in a similar way.

2.4 Computation of penalized likelihoods

How the penalty term $J(\cdot)$ and the covariate in (1.2) and (1.7) specifies a model for lifetime and binary data is introduced in Section 1.1.1 and Section 1.2.1.

The first term of (2.5) 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 (2.5) has an expression

$$\begin{aligned}\zeta(x) &= \sum_{i=1}^{m} d_{\nu,\zeta}\phi_{\nu,\zeta}(x) + \sum_{i=1}^{n} c_{i,\zeta}R_{J,\zeta}(x_i, x) = \boldsymbol{\phi}_\zeta^T(x)\boldsymbol{d}_\zeta + \boldsymbol{\xi}_\zeta^T(x)\boldsymbol{c}_\zeta \\ &= \begin{pmatrix}\boldsymbol{\phi}_\zeta(x) \\ \\ \boldsymbol{\xi}_\zeta(x)\end{pmatrix}^T \begin{pmatrix}\boldsymbol{d}_\zeta \\ \\ \boldsymbol{c}_\zeta\end{pmatrix} \stackrel{Let}{=} \boldsymbol{\psi}_\zeta(x)^T\boldsymbol{b}_\zeta \qquad (2.7)\end{aligned}$$

Substituting (2.7) into (2.5), one calculates the minimizer ζ_β of (2.5) by minimizing

$$\sum_{i=1}^{n}\left\{y_i(\boldsymbol{\phi}_\zeta^T(x_i)\boldsymbol{d}_\zeta + \boldsymbol{\xi}_\zeta^T(x_i)\boldsymbol{c}_\zeta) - \log[1+exp(\boldsymbol{\phi}_\zeta^T(x_i)\boldsymbol{d}_\zeta + \boldsymbol{\xi}_\zeta^T(x_i)\boldsymbol{c}_\zeta)]\right\} + \frac{\beta}{2}\boldsymbol{c}_\zeta^T Q_\zeta \boldsymbol{c}_\zeta, \quad (2.8)$$

with respect to $\boldsymbol{d}_\zeta$ and $\boldsymbol{c}_\zeta$. For a fixed smoothing parameter β, the minimizer ζ_β may be computed via the Newton iteration. Write

$$\tilde{u}_i = -Y_i + \frac{exp[\tilde{\zeta}(x_i)]}{1+exp[\tilde{\zeta}(x_i)]} = -Y_i + \tilde{\mu}(x_i)$$

and

$$\tilde{w}_i = \frac{exp[\tilde{\zeta}(x_i)]}{(1+exp[\tilde{\zeta}(x_i)])^2} = \tilde{v}(x_i).$$

The quadratic approximation of $-Y_i\zeta(x_i) - log(1+exp(\zeta(x_i)))$ at $\tilde{\zeta}(x_i)$ is

$$-\frac{1}{2}\tilde{w}_i\left\{\zeta(x_i) - \tilde{\zeta}(x_i) + \tilde{u}_i/\tilde{w}_i\right\}^2 + C_i \qquad (2.9)$$

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

$$-\frac{1}{n}\sum_{i=1}^{n}\tilde{w}_i(\tilde{Y}_i-\zeta(x_i))^2+\beta J(\zeta) \tag{2.10}$$

where $\tilde{Y}_i=\tilde{\zeta}(x_i)-\tilde{u}_i/\tilde{w}_i$. The selection of the smoothing parameters can be done through an outer-loop optimization of a cross-validation score, see Section 7.2 of Gu [2002].

The first term of (2.6) 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. Let $\{\phi_\nu\}_{\nu=1}^m$ be a basis of $\mathcal{N}_J$. By definition, a function in $\mathcal{H}_N$ has an expression

$$\begin{aligned}\eta(x) &= \sum_{i=1}^{m}d_{\nu,\eta}\phi_{\nu,\eta}(x)+\sum_{i=1}^{n}c_{i,\eta}R_{J,\eta}(x_i,x)=\boldsymbol{\phi}_\eta^T(x)\boldsymbol{d}_\eta+\boldsymbol{\xi}_\eta^T(x)\boldsymbol{c}_\eta \\ &= \begin{pmatrix}\boldsymbol{\phi}_\eta(x)\\ \boldsymbol{\xi}_\eta(x)\end{pmatrix}^T\begin{pmatrix}\boldsymbol{d}_\eta\\ \boldsymbol{c}_\eta\end{pmatrix}\overset{Let}{=}\boldsymbol{\psi}_\eta(x)^T\boldsymbol{b}_\eta\end{aligned} \tag{2.11}$$

where $\boldsymbol{\phi}_\eta$ and $\boldsymbol{\xi}_\eta$ are vectors of functions and $\boldsymbol{d}_\eta$ and $\boldsymbol{c}_\eta$ are vectors of coefficients. Substituting (2.11) into (2.6), one calculates the minimizer η_λ of (2.6) in $\mathcal{H}_N$ by minimizing

$$-\frac{1}{n}\sum_{i=1}^{n}y_i\delta_i(\boldsymbol{\phi}_\eta^T(x_i)\boldsymbol{d}_\eta+\boldsymbol{\xi}_\eta^T(x_i)\boldsymbol{c}_\eta)+\frac{1}{n}\sum_{i=1}^{n}y_i\int_{Z_i}^{X_i}\exp(\boldsymbol{\phi}_\eta^T(x_i)\boldsymbol{d}_\eta+\boldsymbol{\xi}_\eta^T(x_i)\boldsymbol{c}_\eta)dt+\frac{\lambda}{2}\boldsymbol{c}_\eta^TQ_\eta\boldsymbol{c}_\eta \tag{2.12}$$

with respect to $\boldsymbol{d}_\eta$ and $\boldsymbol{c}_\eta$.

Write $\mu_f(g) = n^{-1}\sum_{i=1}^{n}\int_{Z_i}^{X_i} g(t,U_i)e^{f(t,U_i)}dt$ and $V_f(g,h) = \mu_f(gh)$. The minimization of (2.6) for fixed smoothing parameters can be done through Newton iteration, which updates the coefficients from the current iterate $\tilde{\eta} = \boldsymbol{\phi}_\eta^T\tilde{\boldsymbol{d}}_\eta + \boldsymbol{\xi}_\eta^T\tilde{\boldsymbol{c}}_\eta$ through

$$\begin{pmatrix} V_{\phi,\phi} & V_{\phi,\xi} \\ V_{\xi,\phi} & V_{\xi,\xi}+\lambda Q \end{pmatrix}\begin{pmatrix} \boldsymbol{d}_\eta \\ \boldsymbol{c}_\eta \end{pmatrix} = \begin{pmatrix} S^T\mathbf{1}/n - \mu_\phi + V_{\phi,\eta} \\ R^T\mathbf{1}/n - \mu_\xi + V_{\xi,\eta} \end{pmatrix}, \tag{2.13}$$

where $V_{\phi,\phi} = V_{\tilde{\eta}}(\boldsymbol{\phi},\boldsymbol{\phi}^T)$, $V_{\phi,\xi} = V_{\xi,\phi}^T = V_{\tilde{\eta}}(\boldsymbol{\phi},\boldsymbol{\xi}^T)$, $V_{\xi,\xi} = V_{\tilde{\eta}}(\boldsymbol{\xi},\boldsymbol{\xi}^T)$, $\mu_\phi = \mu_{\tilde{\eta}}(\boldsymbol{\phi})$, $\mu_\xi = \mu_{\tilde{\eta}}(\boldsymbol{\xi})$, $V_{\phi,\eta} = V_{\tilde{\eta}}(\boldsymbol{\phi},\tilde{\eta})$, and $V_{\xi,\eta} = V_{\tilde{\eta}}(\boldsymbol{\xi},\tilde{\eta})$; see, e.g., Section 7.1 of Gu [2002]. The selection of the smoothing parameters can be done through an outer-loop optimization of a cross-validation score derived in Section 7.2 of Gu [2002].

2.5 Observed Information Matrix and Confidence Interval

Point estimation alone is often insufficient in practical applications, as it lacks an assessment of the estimation precision. We now derive confidence intervals for the cure rate and hazard estimates of (2.2). Empirical coverage properties of these intervals will be assessed in Section 2.6.

Lacking the parametric sampling distribution, however, an adequately justified interval estimate is a rarity in nonparametric function estimation. An exception to this is the Bayesian confidence intervals of Wahba [1983], which are derived from the Bayes model.

For the general penalized likelihood $L(\eta) + \frac{\lambda}{2}J(\eta)$, 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 $\widehat{\eta}$ 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 $\widehat{\eta}$. 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].

Unfortunately, the techniques described above can not be easily applied to the cure rate model. The reason is that the penalized likelihoods (2.3) and (2.4) in our EM algorithm involve the latent variable $y's$. Instead of directly dealing with observed likelihood, we introduce the unobserved non-cure status y's and work on the complete likelihood to obtain our minimizer of the likelihood function. Introducing the unobserved y's facilitates the estimation, but they are latent variable and should ideally be integrated out from the complete likelihood when considering the confidence interval for the unknown functions. Hence, the quadratic approximation used in estimation can not be used directly for the approximate Bayesian confidence interval calculation. And integrating out the latent variable is too costly to be practical in reality.

This problem is similar to that in the classical EM algorithm. When the algorithm is used to find the MLE for incomplete data, it does not provide a means of

estimating the information matrix associated with the MLE as in the typical maximum likelihood method since the likelihood is partitioned and incomplete data are introduced. In 1982 Louis Louis [1982] first applied the missing information principle to obtain the observed information matrix. The technique described in Louis [1982] requires calculation of the complete data variance-covariance matrix and the conditional expectation of the square of the complete data score function in addition to the EM steps. A popular alternative to the Louis approach for computing the variance-covariance matrix of the EM estimate is the supplemented EM (SEM) algorithm proposed in Meng and Rubin [1991], with further extensions in Jamshidian and Jennrich [2000] and Segal et al. [1994]. The SEM uses the fact that the rate of convergence of EM is governed by the fractions of missing information. It calculates the increased variability due to missing information through a sequence of the converged EM steps and added it to the complete data variance-covariance matrix. However, the SEM algorithm requires numerical differentiation. Although it works well for low-dimensional parameter estimation, it seems to be quite unstable in our problem where the parameters have very high dimensions. So we derive our observed information matrix based on Louis [1982], which applies the missing information principle: $I_{obs} = I_{full} - I_{full|obs}$

$$I_{obs}(\theta) = E_\theta[B(x, Y, \theta)] - E_\theta[S(x, Y, \theta)S(x, Y, \theta)^T]. \tag{2.14}$$

Here S and B are the gradient vector and the negative of the second derivative matrix of the complete log likelihood respectively; θ is the unknown parameter vector; Y is the missing data. This formulation bypasses the observed likelihood and allows working on the complete likelihood to obtain the observed information matrix. The first term in (2.14) is the conditional expected full data observed information matrix and the second term produces the expected information for the conditional distribution of full data given missing data. Notice that all of these conditional expectations can be computed in the EM algorithm using only S and B, need to be evaluated only at the last iteration of the EM procedure.

In our case, the penalized complete log likelihood is:

$$L(\boldsymbol{y};(\zeta,\eta)) = L_1(\zeta,\boldsymbol{y}) - \frac{n\beta}{2}J(\zeta) + L_2(\eta,\boldsymbol{y}) + \frac{n\lambda}{2}J(\eta) \tag{2.15}$$

where $L_1(\zeta,\boldsymbol{y})$ and $L_2(\eta,\boldsymbol{y})$ are defined in (2.4). And correspondingly,

$$\begin{aligned} S(\boldsymbol{y},\boldsymbol{b}_\zeta,\boldsymbol{b}_\eta) &= \begin{pmatrix} \frac{\partial L}{\partial \boldsymbol{b}_\zeta} \\ \\ \frac{\partial L}{\partial \boldsymbol{b}_\eta} \end{pmatrix} \\ &= \begin{pmatrix} \sum_{i=1}^{n}\left[y_i\boldsymbol{\psi}_\zeta(z_i) - \left(1+exp\left\{\boldsymbol{\psi}_\zeta(z_i)^T\boldsymbol{b}_\zeta\right\}\right)^{-1}\boldsymbol{\psi}_\zeta(z_i)\right] - n\beta Q_\zeta^*\boldsymbol{b}_\zeta \\ \sum_{i=1}^{n}\left[\delta_i\boldsymbol{\psi}_\eta(t_i,x_i) - y_i\int\boldsymbol{\psi}_\eta(t_i,x_i)exp\{\boldsymbol{\psi}_\eta(t_i,x_i)^T\boldsymbol{b}_\eta\}dt\right] - n\lambda Q_\eta^*\boldsymbol{b}_\eta \end{pmatrix} \end{aligned} \tag{2.16}$$

$$
\begin{aligned}
B(\boldsymbol{y}, \boldsymbol{b}_\zeta, \boldsymbol{b}_\eta) &= \begin{pmatrix} -\frac{\partial^2 L}{\partial \boldsymbol{b}_\zeta \partial \boldsymbol{b}_\zeta^T} & 0 \\ 0 & -\frac{\partial L}{\partial \boldsymbol{b}_\eta \partial \boldsymbol{b}_\eta^T} \end{pmatrix} \qquad (2.17) \\
\text{where} & \\
-\frac{\partial^2 L}{\partial \boldsymbol{b}_\zeta \partial \boldsymbol{b}_\zeta^T} &= \sum_{i=1}^{n} \left(1 + exp\left\{\boldsymbol{\psi}_\zeta(z_i)^T \boldsymbol{b}_\zeta\right\}\right)^{-2} \boldsymbol{\psi}_\zeta(z_i) \boldsymbol{\psi}_\zeta(z_i)^T + n\beta Q_\zeta^* \\
-\frac{\partial L}{\partial \boldsymbol{b}_\eta \partial \boldsymbol{b}_\eta^T} &= \sum_{i=1}^{n} y_i \left[\int \boldsymbol{\psi}_\eta(t_i, x_i) \boldsymbol{\psi}_\eta(t_i, x_i)^T exp\{\boldsymbol{\psi}_\eta(t_i, x_i)^T \boldsymbol{b}_\eta\} dt\right] + n\lambda Q_\eta^*
\end{aligned}
$$

where Q_ζ^* and Q_η^* are partitioned matrices of appropriate dimensions with the bottom right positions filled by Q defined in (1.4) and the rest by 0. Plug S and B into (2.14) and take expectation with respect to the missing data $\boldsymbol{y}$ to obtain I_{obs}. The B part is a linear function of $\boldsymbol{y}$, so simply replace $\boldsymbol{y}$ with $E[\boldsymbol{y}]$, which is the converged $\boldsymbol{y}^{(\infty)}$, to

obtain $E[B]$. For SS^T part, since $E[y_i y_j] = E[y_i]E[y_j]$ for $i \neq j$ and $E[y_i]^2 = E[y_i]$, it follows:

$$E\left[S(\boldsymbol{y}, \boldsymbol{b}_\zeta, \boldsymbol{b}_\eta) S^T(\boldsymbol{y}, \boldsymbol{b}_\zeta, \boldsymbol{b}_\eta)\right] = E\left[S(\boldsymbol{y}, \boldsymbol{b}_\zeta, \boldsymbol{b}_\eta)\right] E\left[S(\boldsymbol{y}, \boldsymbol{b}_\zeta, \boldsymbol{b}_\eta)\right]^T + \begin{pmatrix} A_{11} & A_{12} \\ A_{12}^T & A_{22} \end{pmatrix} \tag{2.18}$$

where

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

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

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

$$\times \left(\int \boldsymbol{\psi}_\eta(t_i, x_i)^T exp\{\boldsymbol{\psi}_\eta(t_i, x_i)^T \boldsymbol{b}_\eta\} dt\right)$$

Replace $E[\boldsymbol{y}]$ with $\boldsymbol{y}^{(\infty)}$ to obtain the $E[SS^T]$ part. Thus we have the observed information matrix. Let $\boldsymbol{z}_0$ and $\boldsymbol{x}_0$ be the evaluation points for cure rate and survival components respectively. To obtain $100(1-\alpha)\%$ point-wise confidence intervals for the two evaluation points from the observed information matrix, we have

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

corresponding to the two confidence intervals.

This is essentially the quadratic approximation of the posterior distribution based on the equivalence of the Bayesian model after eliminating the impact of the latent variable y's. So we apply the same computation idea described in Gu and Kim [2002] to produce the point-wise confidence interval. 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.

2.6 Model Selection

The purpose of this section is to develop an empirical model selection tool for detecting negligible terms in the SS ANOVA decompositions of $\zeta(\boldsymbol{z})$ and $\eta(\boldsymbol{x})$. In parametric analysis, likelihood ratio tests or the like serve as the primary tool for the purpose. Under nonparametric settings, the parameter space under null hypothesis is typically infinite dimensional, and consequently the sampling distributions of likelihood ratio statistics are no longer available. Hence one has to look elsewhere. Our diagnostic tools are based on the Kullback-Leibler geometry introduced in Gu [2004].

In general, suppose we try to estimate a function f_0 through the minimization of penalized likelihood $L(f) + \frac{\lambda}{2}J(f)$. Suppose the estimation of f_0 has been done in a space of $\mathcal{H}_1$ and one wants to assess the possibility of reducing the model space to a subspace $\mathcal{H}_2 \subset \mathcal{H}_1$. Let $KL(f_1, f_2)$ be the Kullback-Leibler distance between two estimates f_1 and f_2. Let $\hat{f}$ be the estimate of f_0 in $\mathcal{H}_1$. Let $\tilde{f}$ be the Kullback-

Leibler projection of $\hat{f}$ in $\mathcal{H}_2$, that is, the minimizer of $\mathrm{KL}(\hat{f}, f)$ for $f \in \mathcal{H}_2$. Let f_c be the estimate from the constant model. Under many different scenarios, including the nonparametric regression, density estimation and hazard estimation problems considered in Gu [2004], one can derive the following triangle equality:

$$KL(\widehat{f}, f_c) = KL(\widehat{f}, \widetilde{f}) + KL(\widetilde{f}, f_c),$$

where $KL(\widehat{f}, f_c)$ is the "total entropy" and $KL(\widetilde{f}, f_c)$ is the "preserved entropy" by the subspace $\mathcal{H}_2$. $KL(\widehat{f}, f_c)$ quantifies the total amount of variation explained. The ratio $1 - \rho = KL(\widetilde{f}, f_c)/KL(\widehat{f}, f_c)$ is very much like an R^2 statistics. It depicts how much of the variation sits in the subspace $\mathcal{H}_2$. On the other hand, ρ represents the proportion of additional variation explained by the full model space $\mathcal{H}_1$ given the variation explained by the subspace $\mathcal{H}_2$. Thus ρ can be used to identify negligible terms. A small value of ρ indicates the lack of necessity for the extra complexity beyond the parameter space under reduced model.

For our cure rate model, consider the complete log likelihood evaluated at the converged "$y's$". We can define the Kullback-Leibler distance separately for the ζ and η parts:

$$KL(\zeta_1,\zeta_2) = \frac{1}{n}\sum_{i=1}^{n}\{\frac{exp\{\zeta_1(z_i)\}}{1+exp\{\zeta_1(z_i)\}}[\zeta_1(z_i)-\zeta_2(z_i)] \tag{2.19}$$

$$- [log(1+\zeta_1(z_i)) - log(1+\zeta_2(z_i))]\}$$

$$KL(\eta_1,\eta_2) = \frac{1}{n}\sum_{i=1}^{n}\int_0^{t_i}\{e^{\eta_1(t,x_i)}[\eta_1(t,x_i)-\eta_2(t,x_i)] - [e^{\eta_1(t,x_i)} - e^{\eta_2(t,x_i)}]\}. \tag{2.20}$$

Then we can derive the corresponding triangle equality for both Kullback-Leibler distance.

$$KL(\widehat{\zeta},\zeta_c) = KL(\widehat{\zeta},\widetilde{\zeta}) + KL(\widetilde{\zeta},\zeta_c) \tag{2.21}$$

$$KL(\widehat{\eta},\eta_c) = KL(\widehat{\eta},\widetilde{\eta}) + KL(\widetilde{\eta},\eta_c). \tag{2.22}$$

where the derivation of the first equality is in Section 2.9 and the second equality is directly available in Gu [2004]. We can use the ratios $KL(\widetilde{\zeta},\widehat{\zeta})/KL(\widehat{\zeta},\zeta_c)$ and $KL(\widetilde{\eta},\widehat{\eta})/KL(\widehat{\eta},\eta_c)$ to identify negligible terms in the SS ANOVA decompositions of ζ and η.

2.7 Simulation Study

2.7.1 Estimation

We now present some simulations to evaluate the empirical performance of the proposed method in both estimation and 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 for this simulation study can be written as:

$$S_{pop}(t, z, x) = 1 - \pi(z) + \pi(z)[S_0(t)]^{r(x)}$$

where $\pi(z)$ is the cure rate function, $S_0(t)$ is the baseline survival function, and $r(x)$ is the relative risk. Let $h_0(t)$ be the baseline hazard function corresponding to $S_0(t)$. Note we have a proportional structure for the hazard function $h(t, x) = h_0(t)r(x)$.

We consider three test cure rate functions and four different test hazard functions. Thus, we have $3 * 4 = 12$ different test function settings. We also repeat every setting with sample size $n = 400$ and $n = 800$. In addition, we have one reference setting with sample size $n = 400$ that can be described by a parametric model. So in total we have $3 * 4 * 2 + 1 = 25$ 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. To assess the performance of our nonparametric estmation method under a parametric true model, we also considered a simple cure rate function $\pi_4(z) = logit^{-1}[8(z+.045)]$.

For the survival component, the four test functions all have one continuous covariate x. The four test hazard functions functions are:

$$h_1(t,x) = \frac{2.5t^{1.5}}{[1+0.5\sin(2\pi x)]^{2.5}}$$

$$h_2(t,x) = 3.5t^{2.5}(20(x+0.5)^2+0.01)^{1.4}$$

$$h_3(t,x) = \frac{2.5t^{1.5}}{[1+0.5sin(\pi(x+1.4))]^{2.5}}$$

$$h_4(t,x) = \frac{2.5t^{1.5}}{[1+0.5sin(\pi(x+0.6))]^{2.5}}$$

The hazard rate is a function of time and the covariate and thus takes form of a surface. Figure 2.1 shows the four test log hazard surfaces. Note that all the hazard

functions have additive logarithms and belong to the Weibull distribution family with scale parameters depending on the covariate. To assess the performance of our nonparametric estimation method under a parametric true model, we also considered a simple hazard function $h_5(t,x) = e^{-x} = \frac{1t^0}{[e^x]^1}$; see Figure A.1 for the plots of this particular setting.

For settings of sample size $n = 400$, the covariates z and x were generated on a grid of 20 equally spaced values over the range $[-0.4, 0.4]$. Then 20 samples are generated at each covariate point: 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 test hazard functions; 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 only difference is that they are generated on a grid of 32 equally spaced values for z and x over the range $[-0.4, 0.4]$, then 25 samples are generated at each covariate point. The censoring times were generated from Weibull distributions with the parameters chosen in a way so that the observed censoring rate for the three settings 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 log hazard $\eta(t, x_{test})$ with x fixed at a randomly picked evaluation point x_{test}, and t on a grid of size 100 equally spaced on IQR of t at $x = x_{test}$,

and for log hazard $\eta(t_{test}, x)$ with t fixed at a randomly picked evaluation point t_{test} within the range of IQR of t, and x on a grid of size 100 equally spaced on $[-0.4, 0.4]$.

The coverage results in all simulation settings are very similar. We present four settings here and the rest are in the Appendix. Figure 2.2 shows simulation results for test functions $\pi_2(z)$, $h_1(t, x)$ and sample size $n = 800$. The frames on the left show the point-wise coverages of logit cure rate and log hazard against time and covariate. Superimposed in the left frames are the magnitude of the curvature $|logit(\pi''(z))|$ and the histograms showing local sample sizes. Local sample sizes are computed as the average frequencies of 100 replicates in the ranges of $[t_{test} - 0.1, t_{test} + 0.1]$ and $[x_{test} - \frac{0.8}{38}, x_{test} + \frac{0.8}{38}]$. Plotted in the right frames are the true test functions (dash-dotted), the averages of point-wise function estimates (solid), the averages of point-wise 95% CIs (dashed), and the empirical 2.5% and 97.5% percentiles of point-wise function estimates (dotted). Figure 2.3 shows the same simulation results for test functions $\pi_2(z)$, $h_1(t, x)$ and sample size $n = 800$.

As seen in Figure 2.2 and Figure 2.3, the mean function estimates of 100 replicates are very close to the true functions. At the tails, the mean estimates are slightly deviated from the test functions. This is not surprising for nonparametric estimation because information from data at the tails of function are diminishing.

We noticed that the point-wise coverage is very close to the nominal level 0.95. The means of the estimated confidence interval are very close to empirical 2.5% and 97.5% percentile of the 100 function estimates. Note that the band is connected point-wise intervals, with no simultaneous coverage property intended. The widths

of the intervals appear to be of the proper magnitude, and it is reassuring to see the widening of the intervals towards the ends of the axis where information from the data is vanishing. However, 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, lower coverage appears to roughly track high curvature. For example, see the first row of Figure 2.2. 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. Again, this is specially illustrated at some boundaries of function domains, for example, the last row of Figure 2.3.

Figure 2.4 and Figure 2.5 show simulation results of sample sizes $n = 400$ with the same settings as Figure 2.2 and Figure 2.3. 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 coverages do not seem to be affected much by sample sizes.

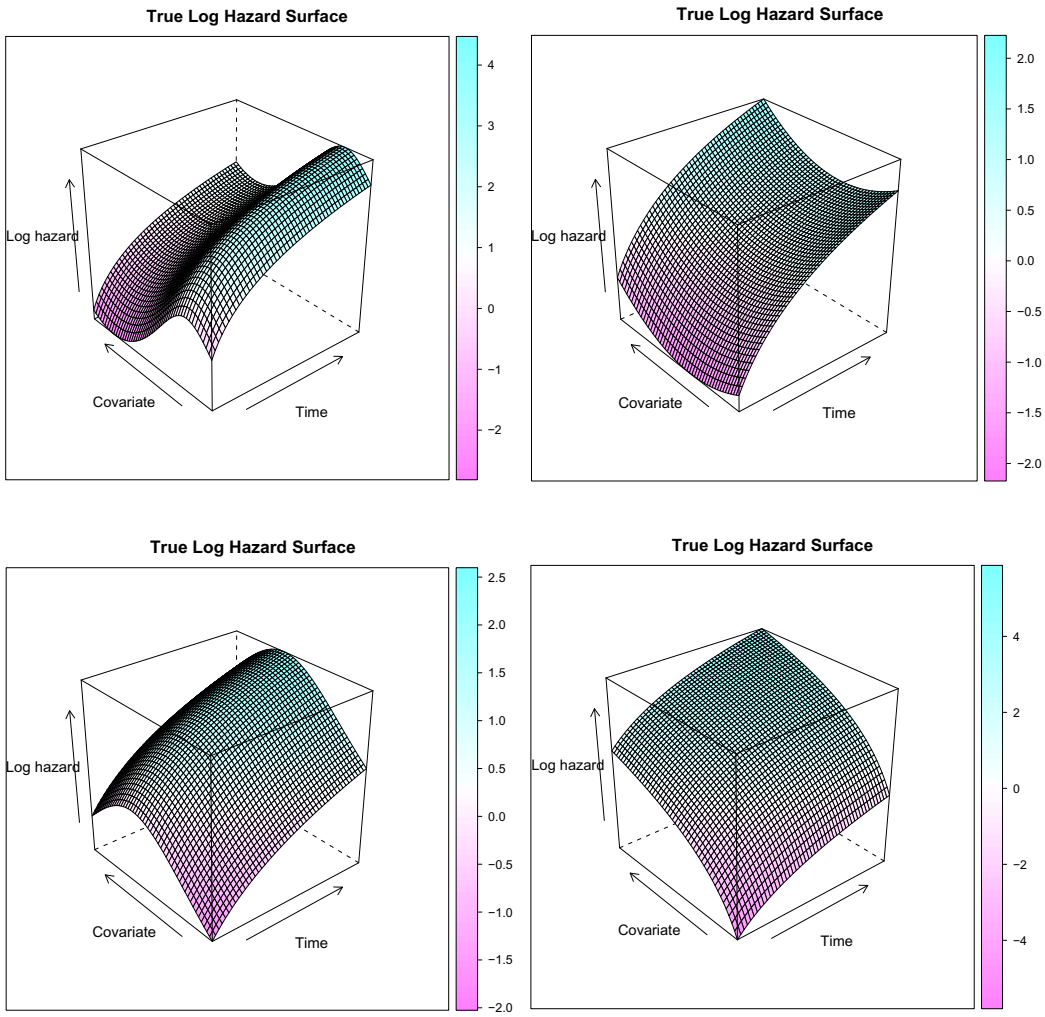

Fig. 2.1. The four test hazard surfaces.

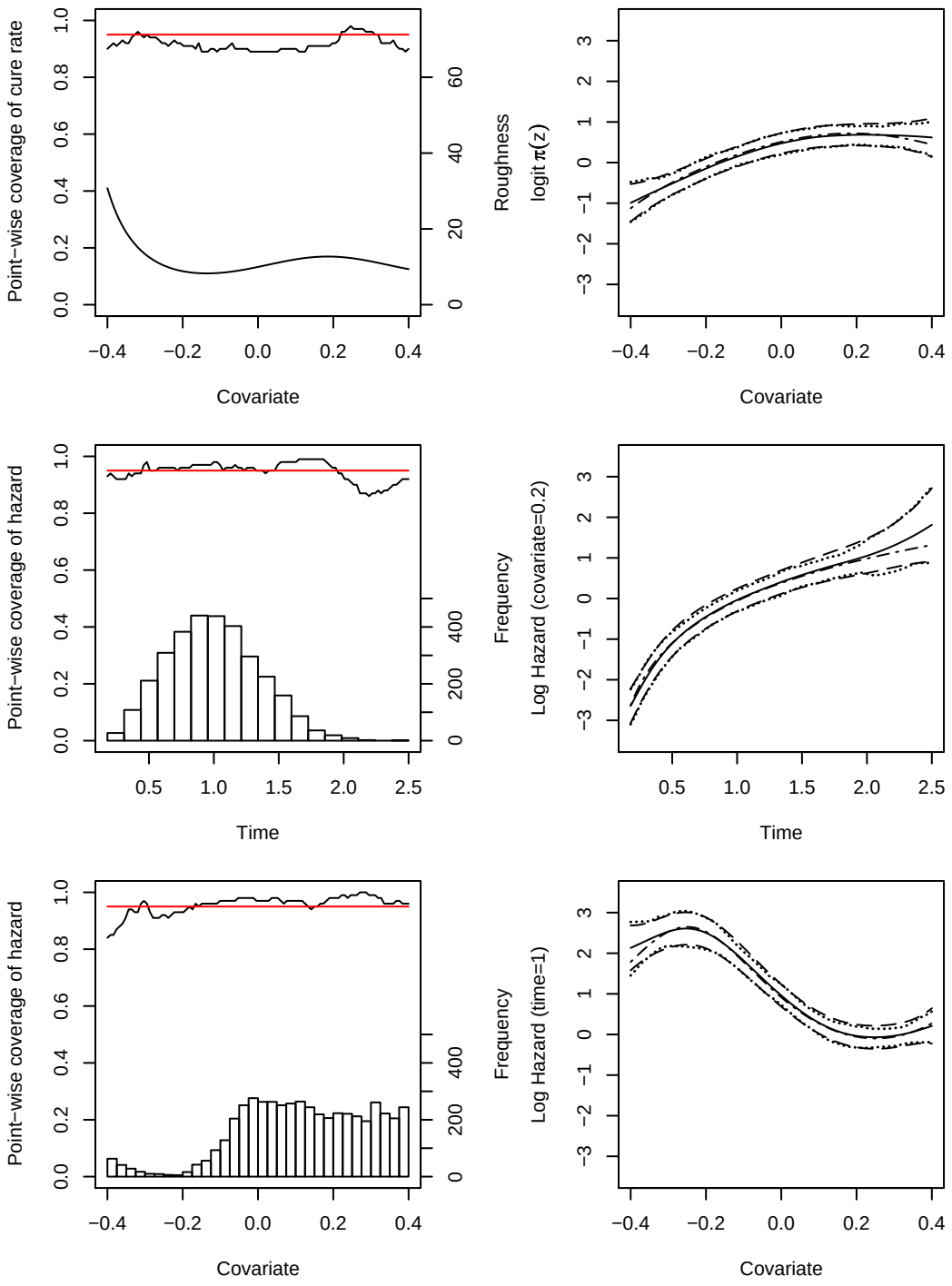

Fig. 2.2. Simulation Results for Test Functions $\pi_2(z)$, $h_1(t,x)$ and $n = 800$. Left column: Point-wise coverages (top black lines). Superimposed are nominal coverage (red lines) and scaled $|logit(\pi''(z))|$ (bottom black line) or histogram of data. Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates.

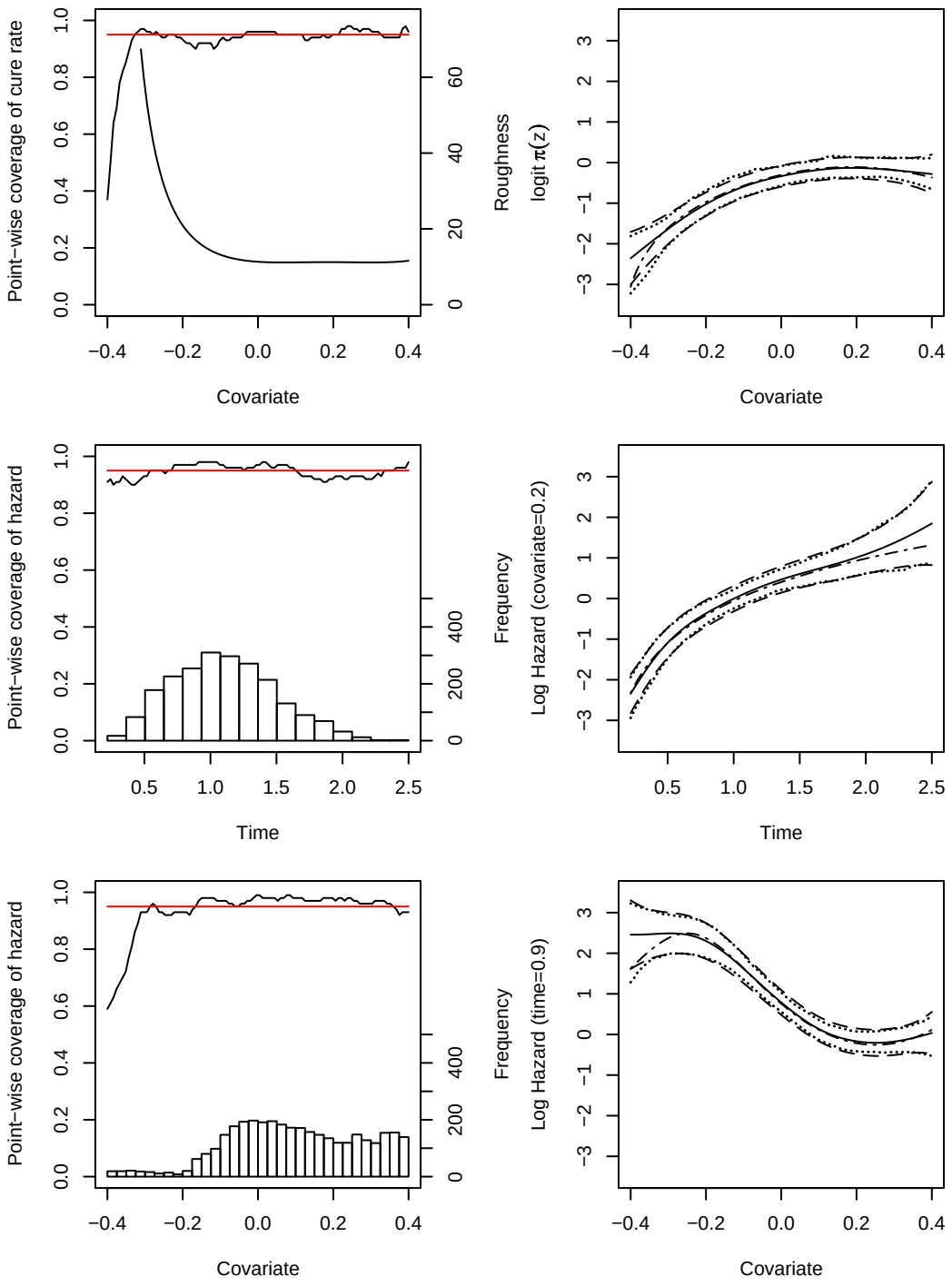

Fig. 2.3. Simulation Results for Test Functions $\pi_3(z)$, $h_1(t,x)$ and $n = 800$. Left column: Point-wise coverages (top black lines). Superimposed are nominal coverage (red lines) and scaled $|logit(\pi''(z))|$ (bottom black line) or histogram of data. Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates.

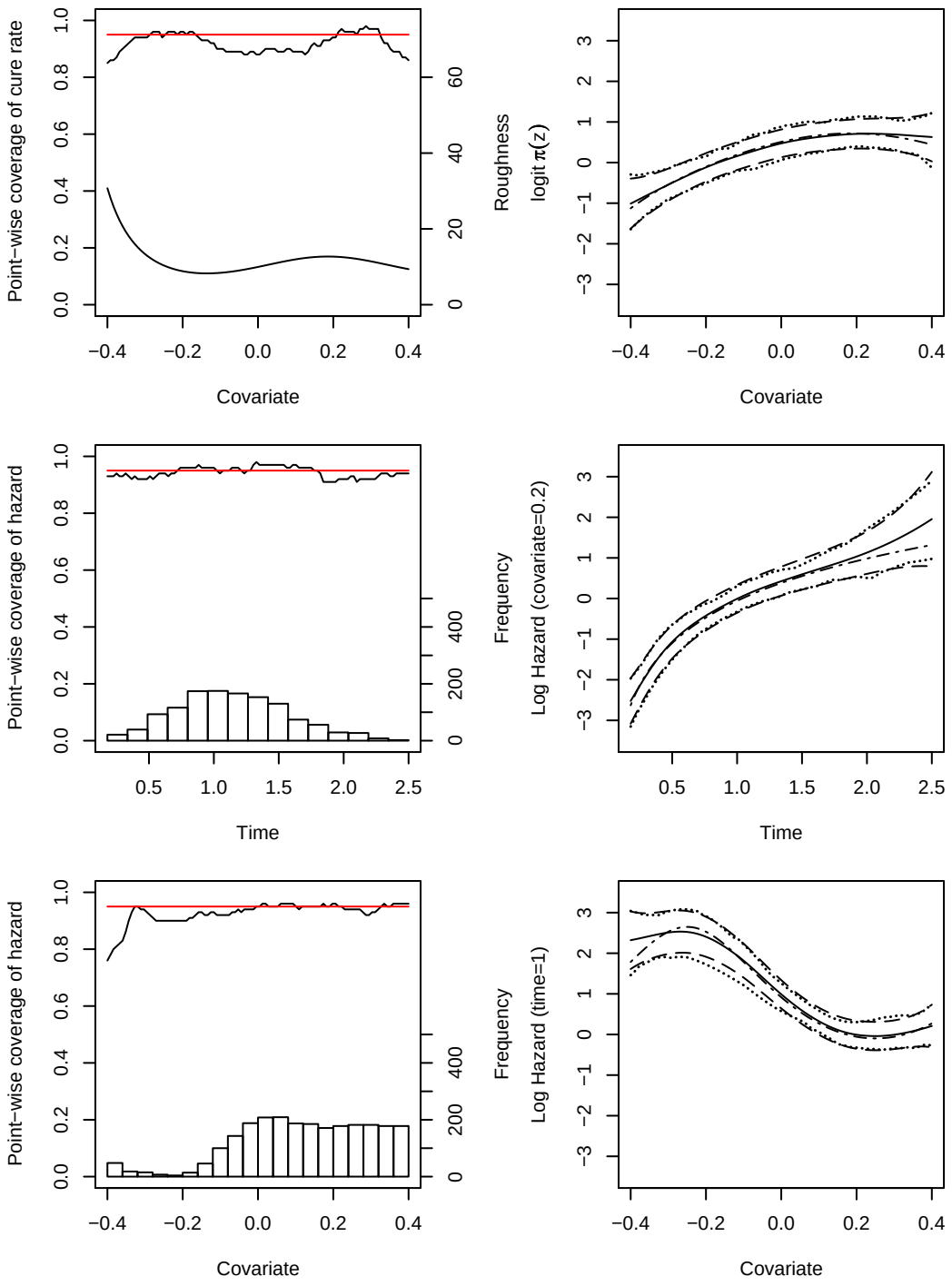

Fig. 2.4. Simulation Results for Test Functions $\pi_2(z)$, $h_1(t,x)$ and $n = 400$. Left column: Point-wise coverages (top black lines). Superimposed are nominal coverage (red lines) and scaled $|logit(\pi''(z))|$ (bottom black line) or histogram of data. Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates.

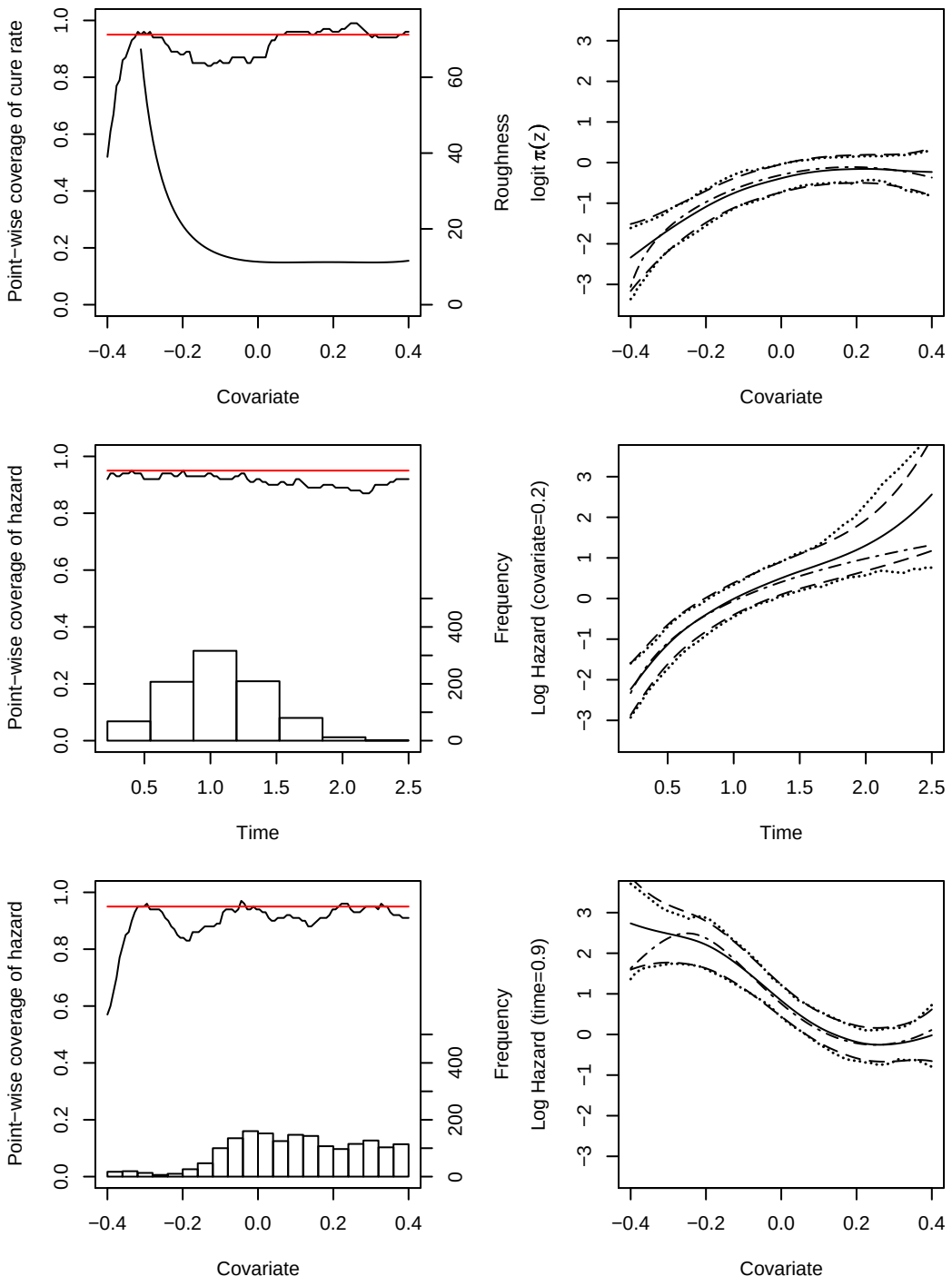

Fig. 2.5. Simulation Results for Test Functions $\pi_3(z)$, $h_1(t,x)$ and $n = 400$. Left column: Point-wise coverages (top black lines). Superimposed are nominal coverage (red lines) and scaled $|logit(\pi''(z))|$ (bottom black line) or histogram of data. Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates.

2.7.2 Model Selection

We consider three settings to assess the model selection tool. We use 0.05 as the cutoff value for ρ, which corresponds to 5% of structure loss measured by the Kullback-Leibler distance if reduced model is adopted. All settings are iterated for 100 times.

In the first setting, we introduce a categorical variable x_c with 2 levels. The true functions are:

$$\pi(z) = -0.0278 + 0.7\sin(2(z + 0.6))$$

$$h(t, x, x_c = 1) = h(t, x, x_c = 0) = \frac{2.5t^{1.5}}{[1 + 0.5\sin(2\pi x)]^{2.5}}.$$

We want to evaluate the model selection tool for the hazard component. Clearly, the true model for log hazard $\eta(t, (x, x_c))$ is additive with only main effects of time and continuous covariate x, or (t, x) using short-hand notation. We consider two scenarios: models (t, x) vs. $(t, x, x_c, x * x_c)$ and models (t) vs. (t, x). The percentage of correct selection are 94% and 100% respectively. See Figure 2.6.

The second setting still focuses on model selection for hazard part. The true cure rate probability function $\pi(z)$ is the same as in the first setting but the true hazard function is:

$$h(t, x, x_c = 1) = 2.5t^{1.5}, \quad h(t, x, x_c = 0) = \frac{2.5t^{1.5}}{2^{2.5}} = 0.44t^{1.5}.$$

Thus the true model for log hazard η is (t, x_c). We again consider two scenarios: models (t, x_c) vs. $(t, x, x_c, x * x_c)$ and models (t) vs. (t, x_c). The percentage of correct selection are 94% and 100% respectively. See Figure 2.7.

The third setting focuses on model selection for cure rate component and introduces an additional categorical variable with 2 levels. The true functions are:

$$\pi(z, z_c = 1) = -0.0278 + 0.7 \sin(2(z + 0.6))$$

$$\pi(z, z_c = 0) = logit^{-1}\{logit(\pi(z, z_c = 1)) + 3.2\}$$

$$h(t, x) = \frac{2.5t^{1.5}}{[1 + 0.5 \sin(2\pi x)]^{2.5}}.$$

Thus the true model for function ζ is (z, z_c). We consider three scenarios: models (z, z_c) vs. $(z, z_c, z * z_c)$, models (z) vs. (z, z_c) and models (z_c) vs. (z, z_c). The percentage of correct selection are 96%, 100% and 97% respectively. See Figure 2.8.

Table 2.1 summarizes the model selection results in terms of proportions of under-fit where some true effect(s) is not selected, correct-fit, and over-fit where some noise effect is selected. Overall the results are very good with high percentages (higher than 90%) of correct fit in all the three settings. And from Figure 2.6-2.8, we can see that even for those few overfits and underfits, the ratios are close to our chosen threshold 0.05. Hence, we will recommend a threshold of 0.05 to use in practice.

Table 2.1
Model Selection Simulation Results. 0.05 is used as the cutoff value for the ρ statistics in the simulation study.

Setting	Function	True Model	Proportion of		
			Under-fit	Correct-fit	Over-fit
I	η	(t, x)	0%	94%	6%
II	η	(t, x_c)	0%	94%	6%
III	ζ	(z, z_c)	3%	93%	4%

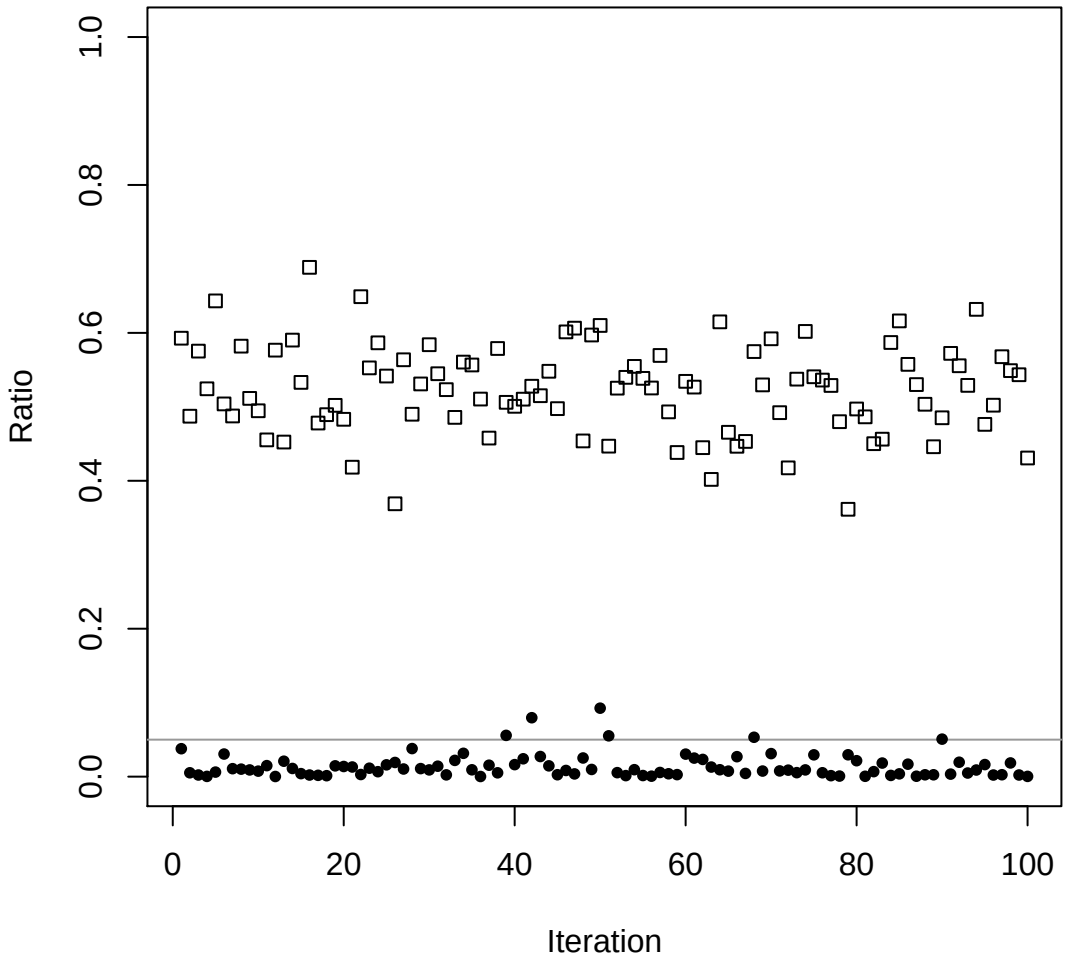

Fig. 2.6. Model selection simulation setting 1. Circle: models (t, x) vs. $(t, x, x_c, x * x_c)$. Square: models (t) vs. (t, x)

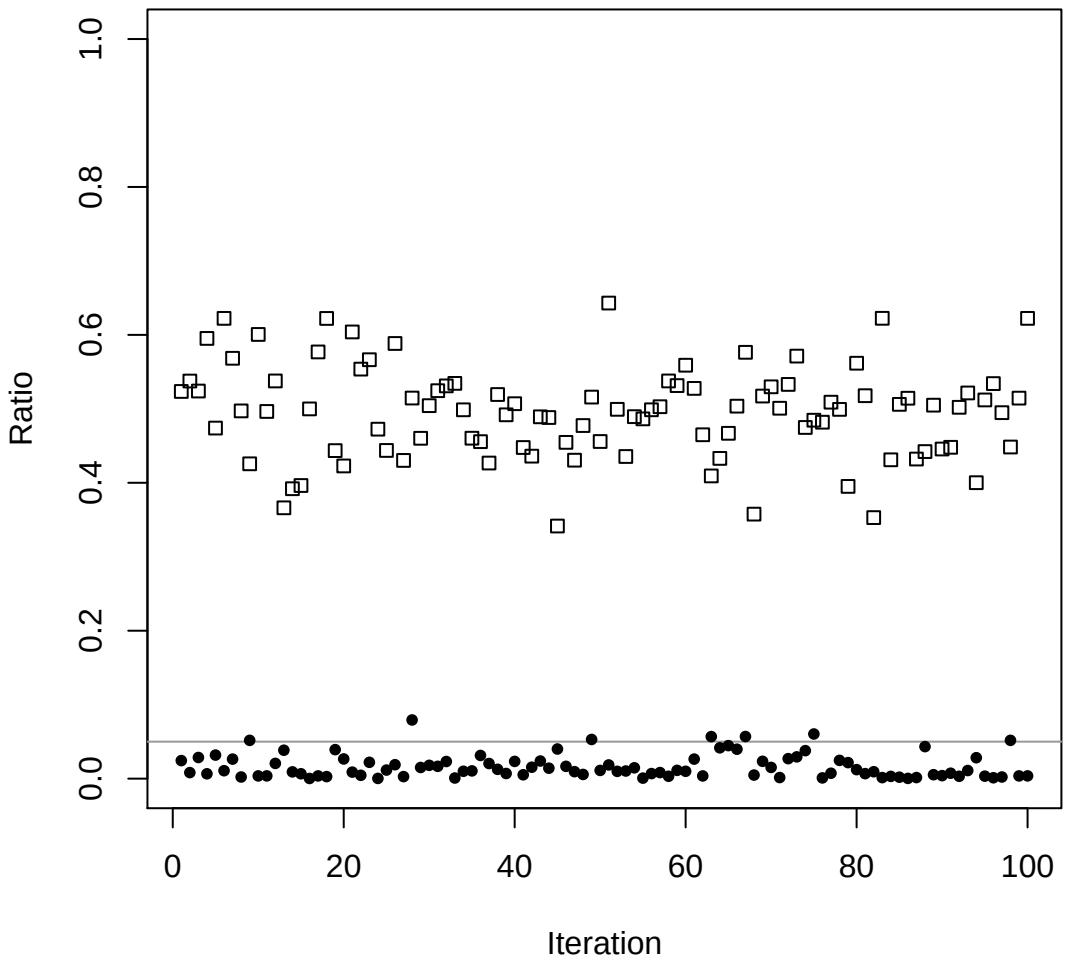

Fig. 2.7. Model selection simulation setting 2. Circle: models (t, x_c) vs. $(t, x, x_c, x * x_c)$. Square: models (t) vs. (t, x)

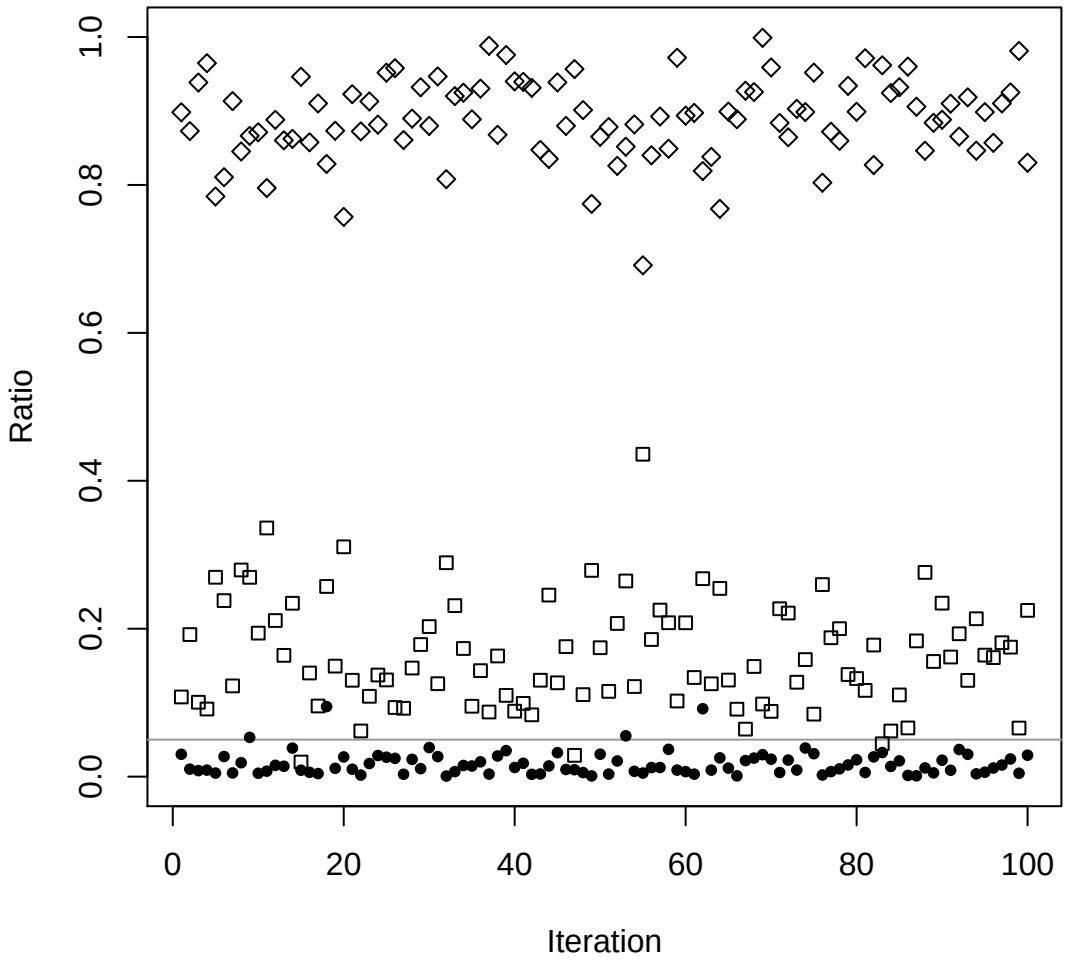

Fig. 2.8. Model selection simulation setting 3. Square: models (z_c) vs. (z, z_c). Diamond: models (z) vs. (z, z_c). Circle: models (z, z_c) vs. $(z, z_c, z * z_c)$

2.8 Melanoma Example

We applied the proposed method to melanoma cancer data from the Surveillance Epidemiology and End Results (SEER) (www.seer.cancer.gov) database, using data released in 2008. Specifically, we looked at males and females diagnosed with melanoma cancer from the all nine registered metropolitan area whose cancer stage was classified as local or regional. To avoid potential confounding related to previous cancer diagnoses, we restricted our sample to the patients whose melanoma cancer was their first cancer diagnosis. Our "failure time" of interest was time from diagnosis of melanoma cancer to death from melanoma cancer. One question of interest was whether survival or cure fractions differed in this data set by gender, tumor size and age. Because of the small number of subjects in some racial categories in this data set, we restricted our attention to the white group. Melanoma cancer is followed by a routine treatment including surgery and radiotherapy for almost most patients but patients with certain medical conditions that prevent such routine treatment. We focused only on patients who received routine treatment. A total of 637 cases in the SEER database met the criteria.

A scatter plot of the data is provided in Figure 2.9. 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 suggest the existence of a subpopulation of cured subjects in the study. Thus a cure rate data analysis is appropriate here.

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 allows interaction between (age, gender, size) for both $\zeta(z)$ and $\eta(t, x)$ such that

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

$$\begin{aligned}\eta(t, (age, gender, size)) = \eta_0 &+ \eta_t(t) + \eta_a(age) + \eta_g(gender) + \eta_s(size) \\ &+ \eta_{ag}(age, gender) + \eta_{as}(age, size) + \eta_{gs}(gender, size) \\ &+ \eta_{ags}(age, gender, size).\end{aligned}$$

Note that besides all the necessary side conditions on the component functions, the exclusion of any time-covariate interactions from the model for η is also to ensure model identifiability.

For cure rate, Figure 2.10 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 cure rate. For male patients, the cure rate increases up to age 60 and then levels off; but for female patients, the 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 cure rate against age for small size tumors seems to be steeper than that for large size tumors. An interesting difference is observed between the male and female groups where large size tumors seem to have a lower cure rate than small size tumors only for the male group.

For the survival component, Figure 2.11 and 2.12 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 2.12. 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 2.11 all show an increasing trend that is very close to be linear. Also notice that the four plots in Figure 2.11 or 2.12 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.

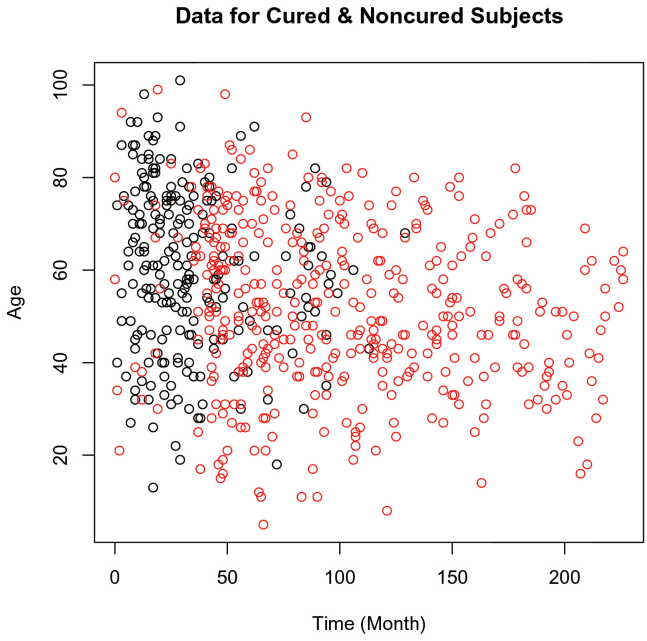

Fig. 2.9. Plot of data. Black circles are observed failures, red circles are observed censoring.

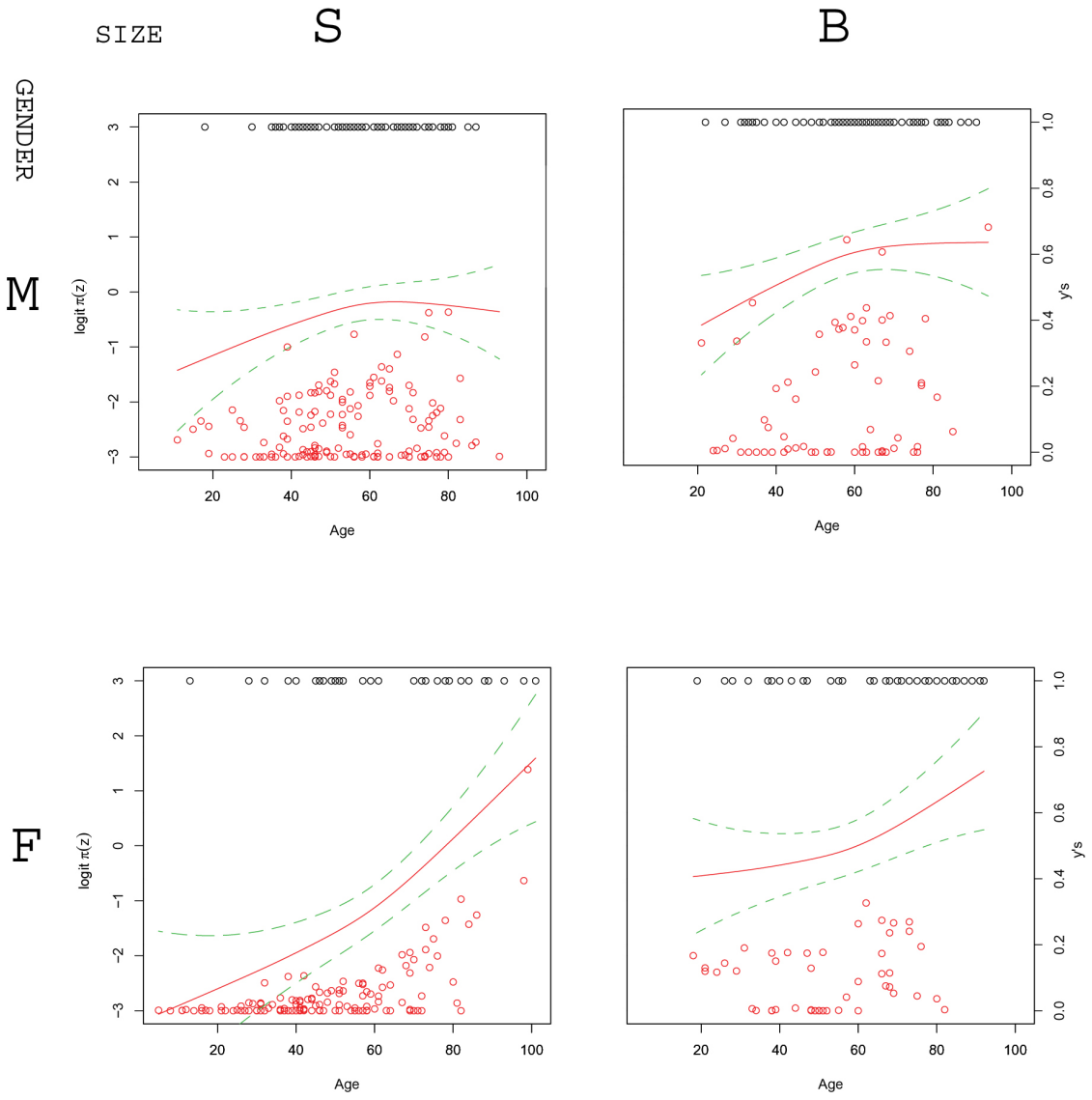

Fig. 2.10. Estimated logit cure rates and their confidence intervals against age. Size: S=small B=big. Gender: M=male F=female. Superimposed are true data points with positions determined by age and converged $y's$. Black circles are observed failures, red circles are observed censoring.

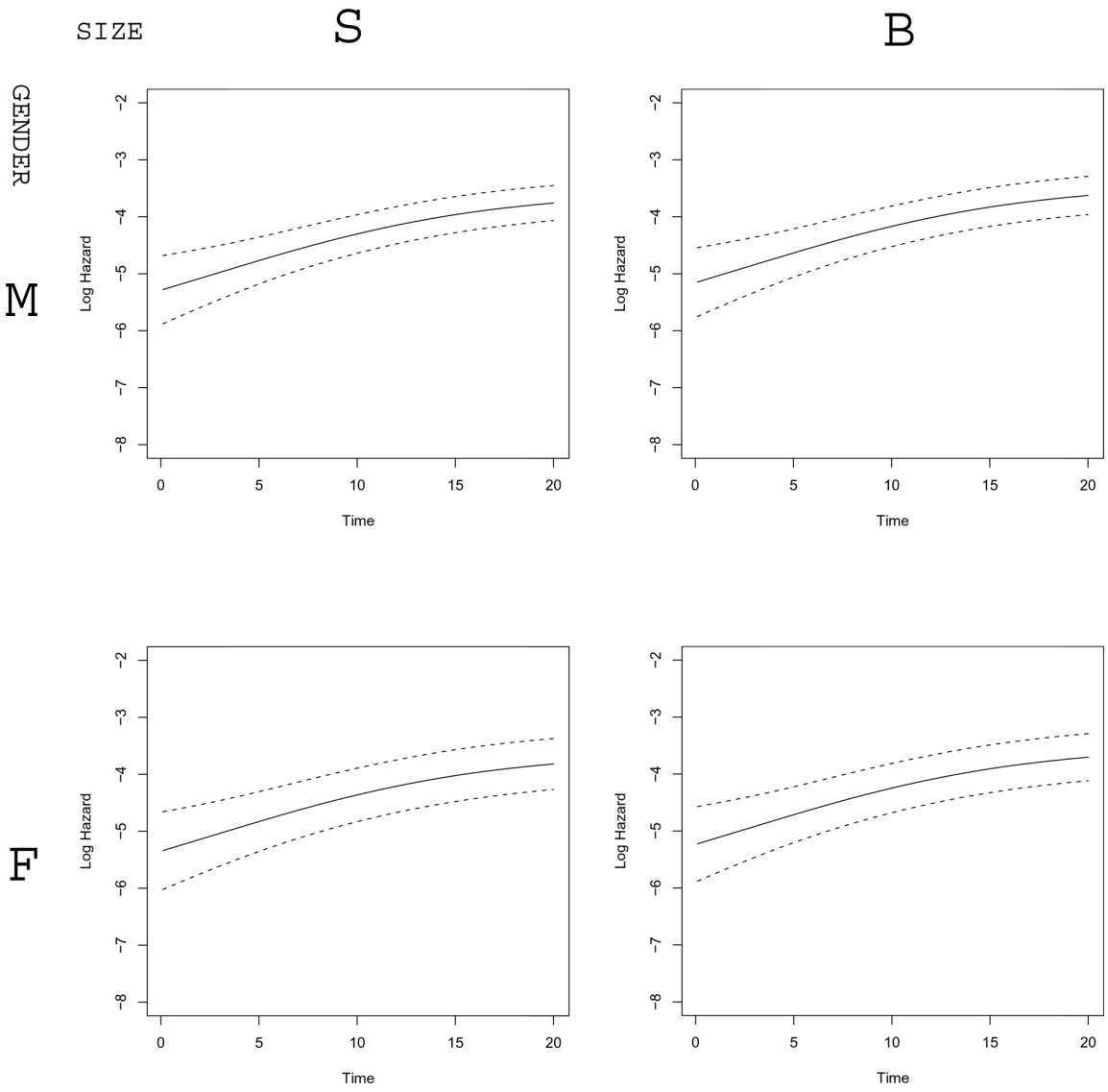

Fig. 2.11. Estimated log hazard and confidence intervals against time at $age = 53$ $years$. Size: S=small B=big. Gender: M=male F=female.

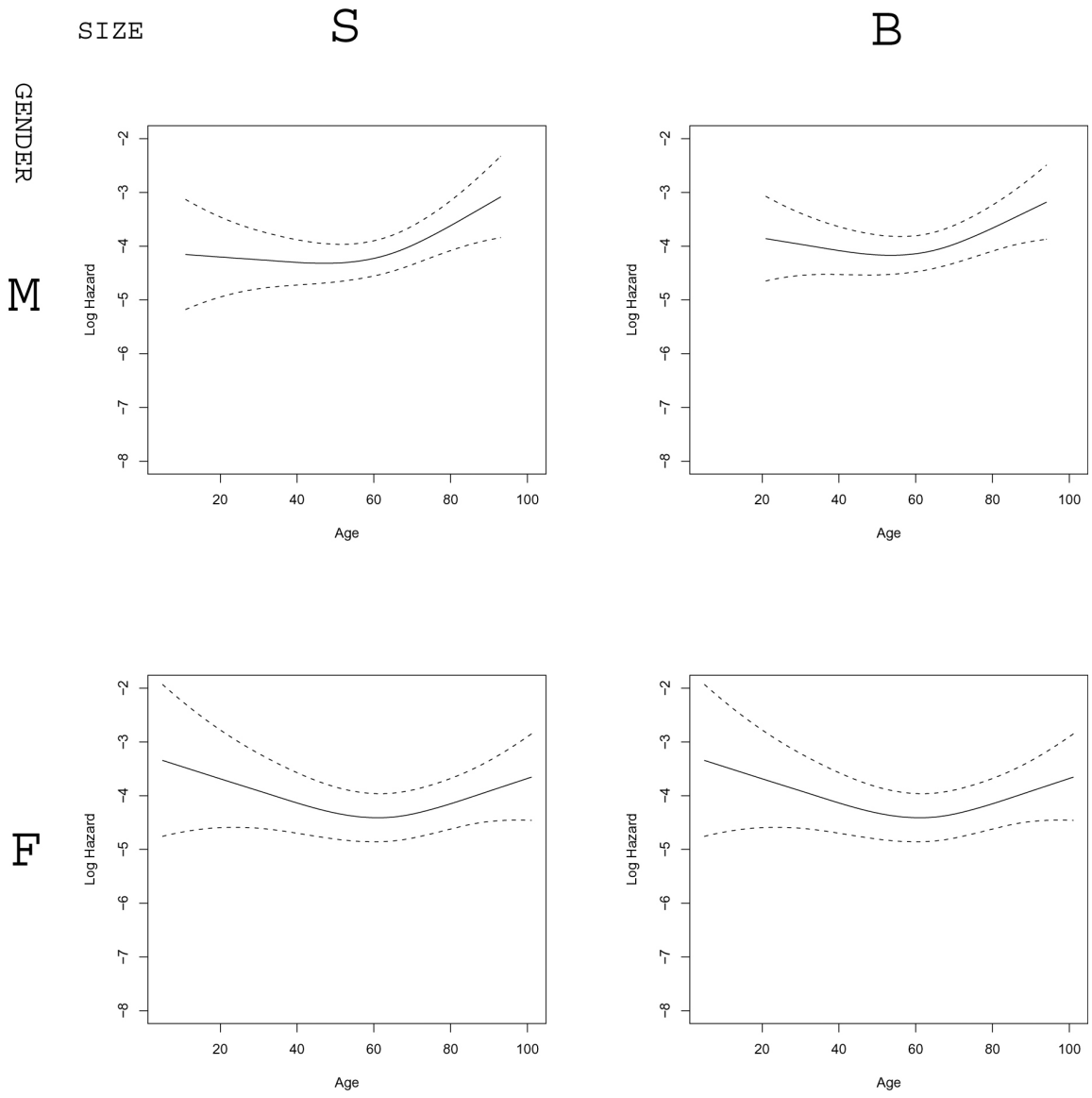

Fig. 2.12. Estimated log hazard and confidence intervals against age at *time* = 10 *months*. Size: S=small B=big. Gender: M=male F=female.

We now apply the model selection tool with a threshold of 0.05 for ρ to this example. Starting with the full models, we summarize the key steps of model selection

for the cure rate component and survival component respectively in Table 2.2 and Table 2.2.

For the cure rate component, the 3-way interaction age*gender*size is not negligible as shown in step one. Consequently, the three 2-way interactions age*gender age*size gender*size should stay in the model. As double checked in step two, 2-way interactions combined with the 3-way interaction are not negligible. Other simpler models also suggest no term could be negligible. So the conclusion is that for the cure rate component age gender size and all their interactions including 2-way and 3-way interaction should stay in the model.

For the survival component, the 3-way interaction age*gender*size is negligible as shown in step one. The combined effect of gender and its associated 2-way interactions age*gender gender*size is negligible as shown in step two. This suggest the variable gender should not be included in the model. Now other than time, only age and size are in the model. As shown in step 3, 2-way interactions age*size is also negligible. Then we update our complete model so it contains time age size. We see size is negligible and only time and age are left in the model. More simpler models suggest no further reduction is possible. So the conclusion is that for the survival component only the main effect time and age should stay in the model.

Table 2.2
Model selection of the cure rate component of melanoma example. ρ is the model selection statistics. A ρ value smaller than 0.05 indicates the reduced model.

Key Model Selection Steps: The Cure Rate Component		
Reduced model terms	ρ	Result
age gender size age*gender age*size gender*size	0.107	age*gender*size in
age gender size	0.153	age*gender age*size gender*size in
gender size	0.768	age in
age size	0.330	size in
age gender	0.153	gender in

Then the model terms for $\boldsymbol{x}$ are (time, age). The model terms for $\boldsymbol{z}$ are (age, gender, size, age*gender, gender*size, age*size, age*gender*size). Thus the final models are:

$$\begin{aligned}\zeta(age, gender, size) =& \zeta_0 + \zeta_a(age) + \zeta_g(gender) + \zeta_s(size) \\ &+ \zeta_{ag}(age, gender) + \zeta_{as}(age, size) + \zeta_{gs}(gender, size) \\ &+ \zeta_{ags}(age, gender, size) \\ \eta(t, (age)) =& \eta_0 + \eta_t(t) + \eta_a(age).\end{aligned}$$

For fitting of the cure rate component, Figure 2.13 shows very similar result as in Figure 2.10. For fitting of the survival component, Figure 2.15 and 2.16 also show similar result as in Figure 2.11 and 2.12 respectively after averaging over gender and size effects.

Table 2.3

Model selection of the survival component of melanoma example. ρ is the model selection statistics. A ρ value smaller than 0.05 indicates the reduced model.

Key Model Selection Steps		
Reduced model terms	ρ	Result
time age gender size age*gender age*size gender*size	0.002	age*gender*size out
time age size age*size	0.046	gender gender*age gender*size out
time age size	0.049	age*size out
Full model terms: time age size		
time age	0.00002	size out
time	0.162	age in

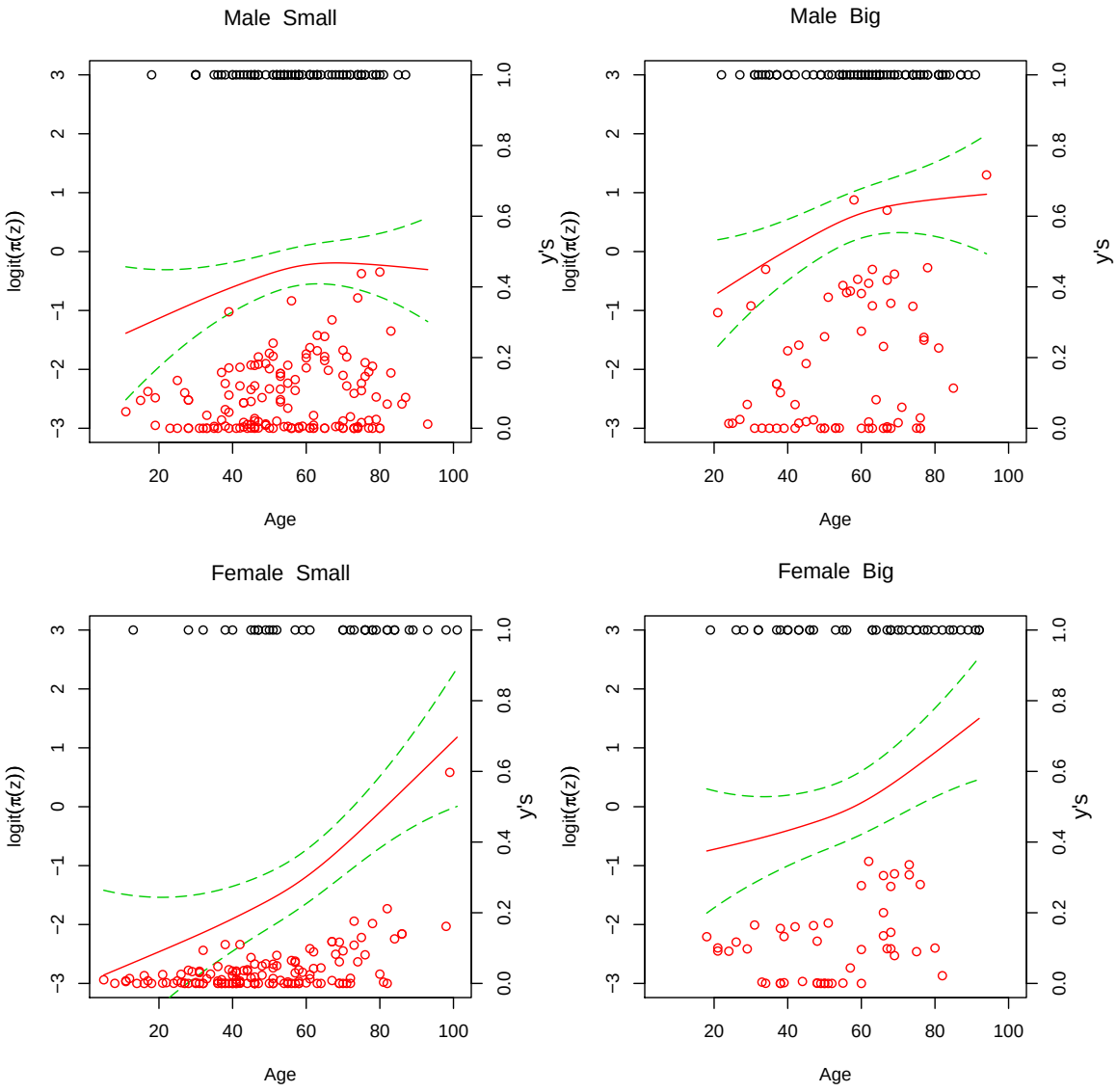

Fig. 2.13. Estimated logit cure rates and their confidence intervals against age. Superimposed are true data points with positions determined by age and converged $y's$. Black circles are observed failures, red circles are observed censoring.

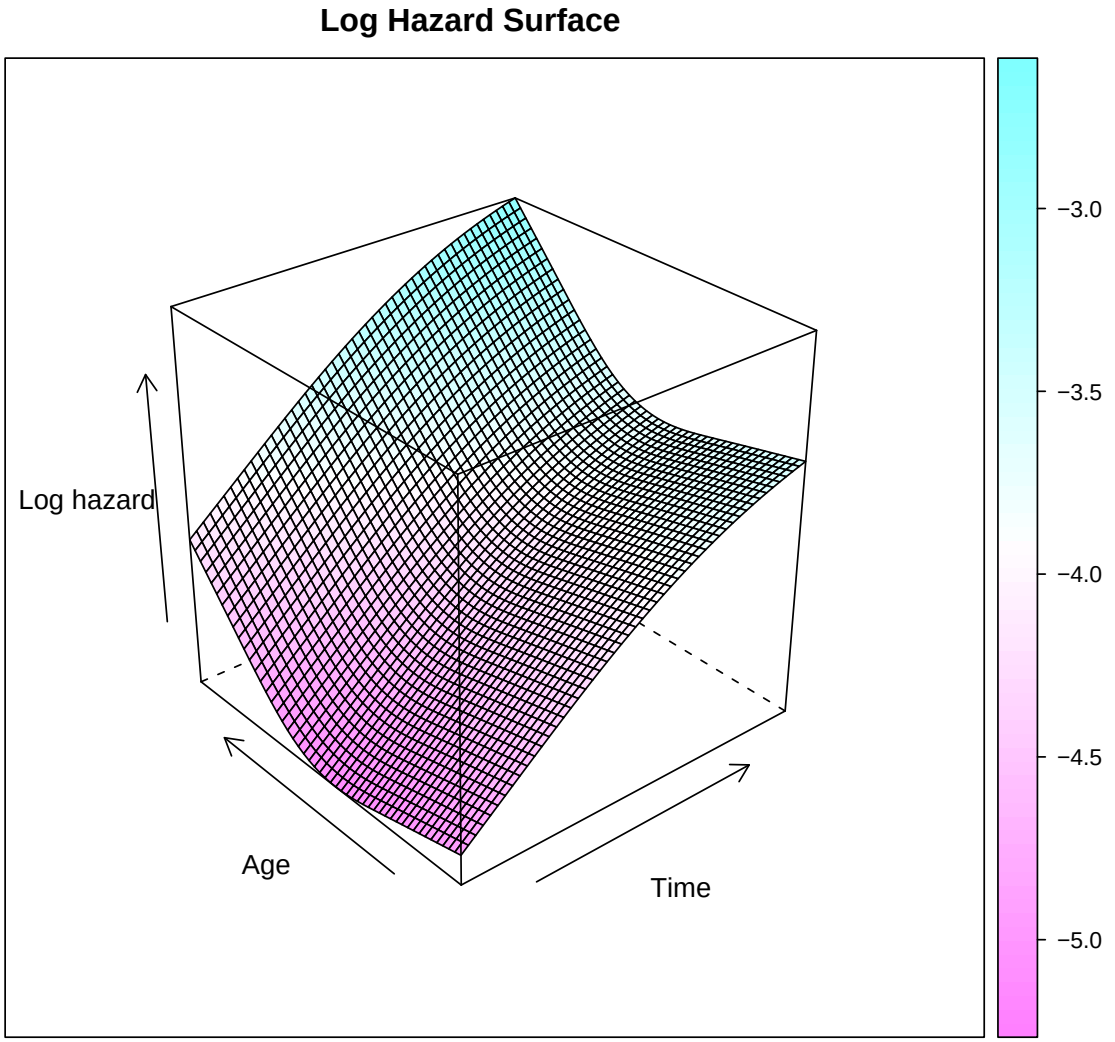

Fig. 2.14. Estimated log hazard surfaces.

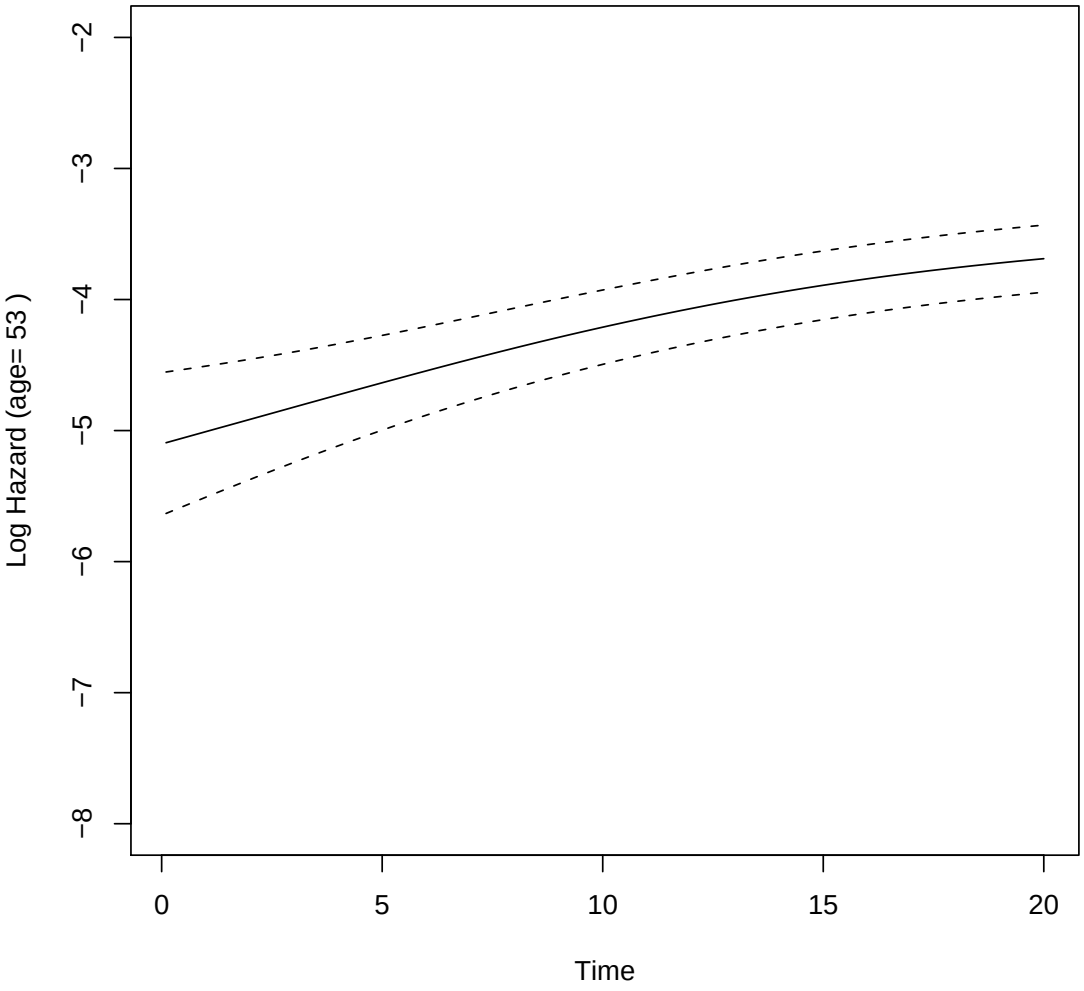

Fig. 2.15. Estimated log hazard and confidence intervals against time at $age = 53\ years$.

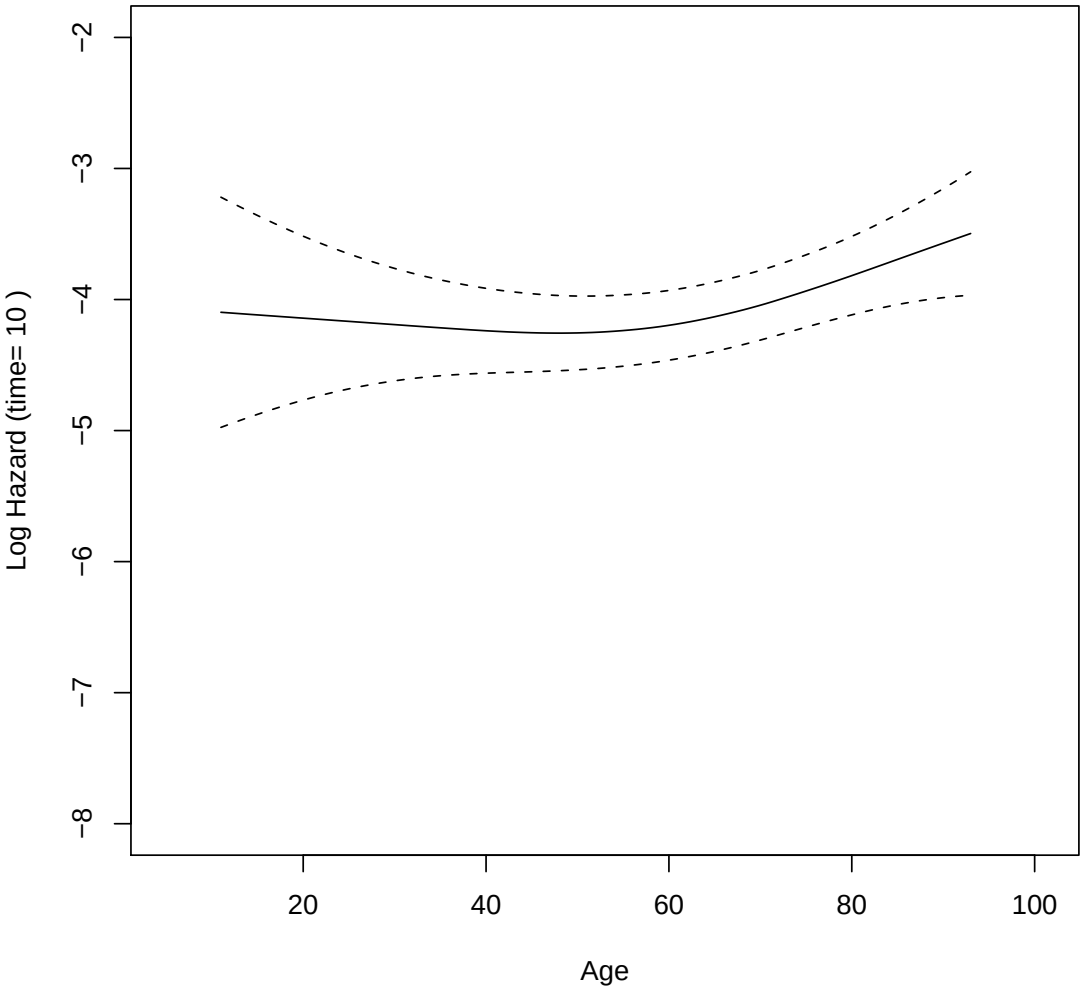

Fig. 2.16. Estimated log hazard and confidence intervals against age at $time = 10$ $months$.

2.9 Proof

2.9.1 Proof of Proposition 2.2.1 (i)

The covariates $\boldsymbol{x}^T = (x_1, x_2, \ldots x_{p_x})$ and $\boldsymbol{z}^T = (z_1, z_2, \ldots z_{p_z})$ can be different, same, or share some common covariates. So there is $\boldsymbol{u}^T = (u_1, u_2, \ldots u_p)$, such that: $\{x_1, x_2, \ldots x_{p_x}\} \subset \{u_1, u_2, \ldots u_p\}$ and$\{z_1, z_2, \ldots z_{p_z}\} \subset \{u_1, u_2, \ldots u_p\}$

We can define:

$$\pi_u(\boldsymbol{u}) = \pi(\boldsymbol{z}) \text{ and } r_u(\boldsymbol{u}) = r(\boldsymbol{x})$$

so,

$$S_{pop}(t, z, x) = S_{pop}(t, \boldsymbol{u}) = 1 - \pi_u(\boldsymbol{u}) + \pi_u(\boldsymbol{u}) S(t, \boldsymbol{u})$$

We need to show that $S_{pop}(t, \boldsymbol{u}) = S^*_{pop}(t, \boldsymbol{u})$ if and only if $\pi(\boldsymbol{u}) = \pi^*(\boldsymbol{u})$ and $S(t, \boldsymbol{u}) = S^*(t, \boldsymbol{u})$. The "if" part is clearly true. To show the "only if" part, suppose that $S_{pop}(t, \boldsymbol{u}) = S^*_{pop}(t, \boldsymbol{u})$. So the ratio

$$\frac{\pi(\boldsymbol{u})}{\pi^*(\boldsymbol{u})} = \frac{1 - S^*(t, \boldsymbol{u})}{1 - S(t, \boldsymbol{u})} \stackrel{let}{=} c(\boldsymbol{u}) \tag{2.23}$$

must be a positive function $c(\boldsymbol{u})$ depending only on $\boldsymbol{u}$ since the left hand side depends only on $\boldsymbol{u}$ but not t. Thus, we have

$$S^*(t, \boldsymbol{u}) = 1 - c(\boldsymbol{u}) + c(\boldsymbol{u}) S(t, \boldsymbol{u}) \tag{2.24}$$

$$\pi^*(\boldsymbol{u}) = \frac{\pi(\boldsymbol{u})}{c(\boldsymbol{u})}. \tag{2.25}$$

If we can find such a function $c(\boldsymbol{u})$ other than 1, then the proof is done. It is clear that any function $c(\boldsymbol{u})$ other than 1 within the range $(max(\pi(\boldsymbol{u})), \frac{1}{1-S(t,\boldsymbol{u})})$, appropriately defines some $\pi(\boldsymbol{u}) \neq \pi^*(\boldsymbol{u})$ and $S(t,\boldsymbol{u}) \neq S^*(t,\boldsymbol{u})$ such that (2.23) holds. So the $S_{pop}(t,z,x)$ is not uniquely represented and not identifiable. □

2.9.2 Proof of Proposition 2.2.1 (ii)

The covariates $\boldsymbol{x}^T = (x_1, x_2, \ldots x_{p_x})$ and $\boldsymbol{z}^T = (z_1, z_2, \ldots z_{p_z})$ can be different, same, or share some common covariates. So there is $\boldsymbol{u}^T = (u_1, u_2, \ldots u_p)$, such that:

$\{x_1, x_2, \ldots x_{p_x}\} \subset \{u_1, u_2, \ldots u_p\}$ and$\{z_1, z_2, \ldots z_{p_z}\} \subset \{u_1, u_2, \ldots u_p\}$

We can define:

$$\pi_u(\boldsymbol{u}) = \pi(\boldsymbol{z}) \text{ and } r_u(\boldsymbol{u}) = r(\boldsymbol{x})$$

so,

$$S_{pop}(t,z,x) = S_{pop}(t,\boldsymbol{u}) = 1 - \pi_u(\boldsymbol{u}) + \pi_u(\boldsymbol{u})[S(t)]^{r_u(\boldsymbol{u})}$$

Since any multiplicative constant can be absorbed into the arbitrary baseline function, we standardize $\boldsymbol{u}$ and $r(\boldsymbol{u})$ such that the space of x includes $\mathbf{0}$ and $r(\mathbf{0}) = \mathbf{1}$.

If the model is not identifiable then there exists functions$(\pi^*(\boldsymbol{u}),\ S^*(t),\ r^*(\boldsymbol{u}))$ not equal to $(\pi(\boldsymbol{u}),\ S(t),\ r(\boldsymbol{u}))$ such that

$$\frac{\pi(\boldsymbol{u})}{\pi^*(\boldsymbol{u})} = \frac{1 - [S^*(t)]^{r^*(\boldsymbol{u})}}{1 - [S(t)]^{r(\boldsymbol{u})}} \stackrel{let}{=} c(\boldsymbol{u}) \tag{2.26}$$

Solving for $\pi^*(\boldsymbol{u})$, $S^*(t)$ and $r^*(\boldsymbol{u})$, we have

$$[S^*(t)]^{r^*(\boldsymbol{u})} = 1 - c(\boldsymbol{u}) + c(\boldsymbol{u})[S(t)]^{r(\boldsymbol{u})}, \tag{2.27}$$

$$\pi^*(\boldsymbol{u}) = \frac{\pi(\boldsymbol{u})}{c(\boldsymbol{u})}. \tag{2.28}$$

At $u = 0$, we have $S^*(t) = 1 - c(0) + c(0)S(t)$, and hence

$$r^*(\boldsymbol{u}) = \frac{log\{1 - c(\boldsymbol{u}) + c(\boldsymbol{u})[S(t)]^{r(\boldsymbol{u})}\}}{log[1 - c(0) + c(0)S(t)]}. \tag{2.29}$$

Let $g = c(\boldsymbol{u})$, then $r(\boldsymbol{u}) = r(c^{-1}(g))$. The first order Taylor series expansion of $\pi^*(\boldsymbol{u})$ at $g = c(0)$ gives

$$\begin{aligned} r^*(\boldsymbol{u}) = & \underbrace{\frac{log\{1 - c(0) + c(0)[S(t)]^{r(c^{-1}(g))}\}}{log[1 - c(0) + c(0)S(t)]}}_{=1} \\ & + \Delta \frac{g^* r(c^{-1}(g^*))[S(t)]^{r(c^{-1}(g^*))-1} \times \frac{r^{(1)}(c^{-1}(g^*))}{c^{(1)}(g^*)} + [S(t)]^{r(c^{-1}(g^*))} - 1}{\{1 - g^* + g^*[S(t)]^{r(c^{-1}(g^*))}\} log[1 - c(0) + c(0)S(t)]} \end{aligned} \tag{2.30}$$

where g^* is between $c(0)$ and $c(0) + \Delta$. For the right hand side of (2.30) not to be a function of $S(t)$ requires $\Delta = 0$. Hence $g = c(x)$ is constant. Thus $r^*(\boldsymbol{u}) = r(\boldsymbol{u})$, and $S_{pop}(t, z, x)$ is uniquely represented and identifiable. □

2.9.3 Proof of 2.21

Substituting in the expression 2.21 with the version associated with subspace $\mathcal{H}_2$, the minimization of 2.21 can be solved via Newton iteration and an outer loop of optimization with respect to the θ_β's. Setting $\zeta = \widetilde{\zeta} + \alpha(\widetilde{\zeta} - \zeta_c)$ in 2.21 for α real, then 2.21 can be written as

$$K(\alpha) = KL(\widehat{\zeta}, \zeta) = \frac{1}{n}\sum_{i=1}^{n}\{\frac{exp\{\widehat{\zeta}(z_i)\}}{1+exp\{\widehat{\zeta}(z_i)\}}[\widehat{\zeta}(z_i) - \widetilde{\zeta}(z_i) - \alpha(\widetilde{\zeta}(z_i) - \zeta_c(z_i))] \\ - [log(1+\widehat{\zeta}(z_i)) - log(1+\widetilde{\zeta}(z_i) - \alpha(\widetilde{\zeta}(z_i) - \zeta_c(z_i)))]\}. \tag{2.31}$$

Differentiate with respect to α, and evaluate at $\alpha = 0$, one has

$$\frac{1}{n}\sum_{i=1}^{n}\{[\frac{exp\{\widetilde{\zeta}(z_i)\}}{1+exp\{\widetilde{\zeta}(z_i)\}} - \frac{exp\{\widehat{\zeta}(z_i)\}}{1+exp\{\widehat{\zeta}(z_i)\}}][\widetilde{\zeta}(z_i) - \zeta_c(z_i)]\} = 0 \tag{2.32}$$

Hence,

$$\frac{1}{n}\sum_{i=1}^{n}\{[\frac{exp\{\widetilde{\zeta}(z_i)\}}{1+exp\{\widetilde{\zeta}(z_i)\}} - \frac{exp\{\widehat{\zeta}(z_i)\}}{1+exp\{\widehat{\zeta}(z_i)\}}][\widetilde{\zeta}(z_i) - \zeta_c(z_i)]\} = 0 \tag{2.33}$$

Then, we have

$$KL(\widehat{\zeta},\widetilde{\zeta}) + KL(\widetilde{\zeta},\zeta_c) \tag{2.34}$$

$$=\frac{1}{n}\sum_{i=1}^{n}\{\frac{exp\{\widehat{\zeta}(z_i)\}}{1+exp\{\widehat{\zeta}(z_i)\}}[\widehat{\zeta}(z_i)-\widetilde{\zeta}(z_i)] - [log(1+\widehat{\zeta}(z_i)) - log(1+\widetilde{\zeta}(z_i))] \tag{2.35}$$

$$+\frac{exp\{\widetilde{\zeta}(z_i)\}}{1+exp\{\widetilde{\zeta}(z_i)\}}[\widetilde{\zeta}(z_i)-\zeta_c(z_i)] - [log(1+\widetilde{\zeta}(z_i)) - log(1+\zeta_c(z_i))]\} \tag{2.36}$$

$$=\frac{1}{n}\sum_{i=1}^{n}\{\frac{exp\{\widehat{\zeta}(z_i)\}}{1+exp\{\widehat{\zeta}(z_i)\}}[\widehat{\zeta}(z_i)-\zeta_c(z_i)] - [log(1+\widehat{\zeta}(z_i)) - log(1+\zeta_c(z_i))]\} \tag{2.37}$$

$$=KL(\widehat{\zeta},\zeta_c) \tag{2.38}$$

□

LIST OF REFERENCES

J. A. Anderson and A. Senthilselvan. Smooth estimates for the hazard function. *J. Roy. Statist. Soc. Ser. B*, 42:322–327, 1980.

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.

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.

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. 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. 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 and Y.-J. Kim. Penalized likelihood regression: General formulation and efficient approximation. *Can. J. Statist.*, 30:619–628, 2002.

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.

M. Jamshidian and R. I. Jennrich. Standard errors for em estimation. *J. Roy. Statist. Soc. Ser. B*, 62(2):257–270, 2000.

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.-S. Li, J. M. G. Taylor, and J. P. Sy. Identifiability of cure models. *Statist. Probab. Lett.*, 54:389–395, 2001.

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. Using EM to obtain asymptotic variance-covariance matrices: The sem algorithm. *J. Amer. Statist. Assoc.*, 86:899–909, 1991.

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.

F. O'Sullivan. Fast computation of fully automated log-density and log-hazard estimators. *SIAM J. Sci. Statist. Comput.*, 9:363–379, 1988a.

F. O'Sullivan. Nonparametric estimation of relative risk using splines and cross-validation. *SIAM J. Sci. Statist. Comput.*, 9:531–542, 1988b.

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.

M. R. Segal, P. Bacchetti, and N. P. Jewell. Variances for maximum penalized likelihood estimates obtained via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 56 (2):345–352, 1994.

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

P. Tai, E. Yu, G. Cserni, G. Vlastos, M. Royce, I. Kunkler, and V. Vinh-Hung. Minimum follow-up time required for the estimation of statistical cure of cancer patients: Verification using data from 42 cancer sites in the seer database. *BMC Cancer*, 5(48), 2005.

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

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.

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, 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.

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. SIMULATION

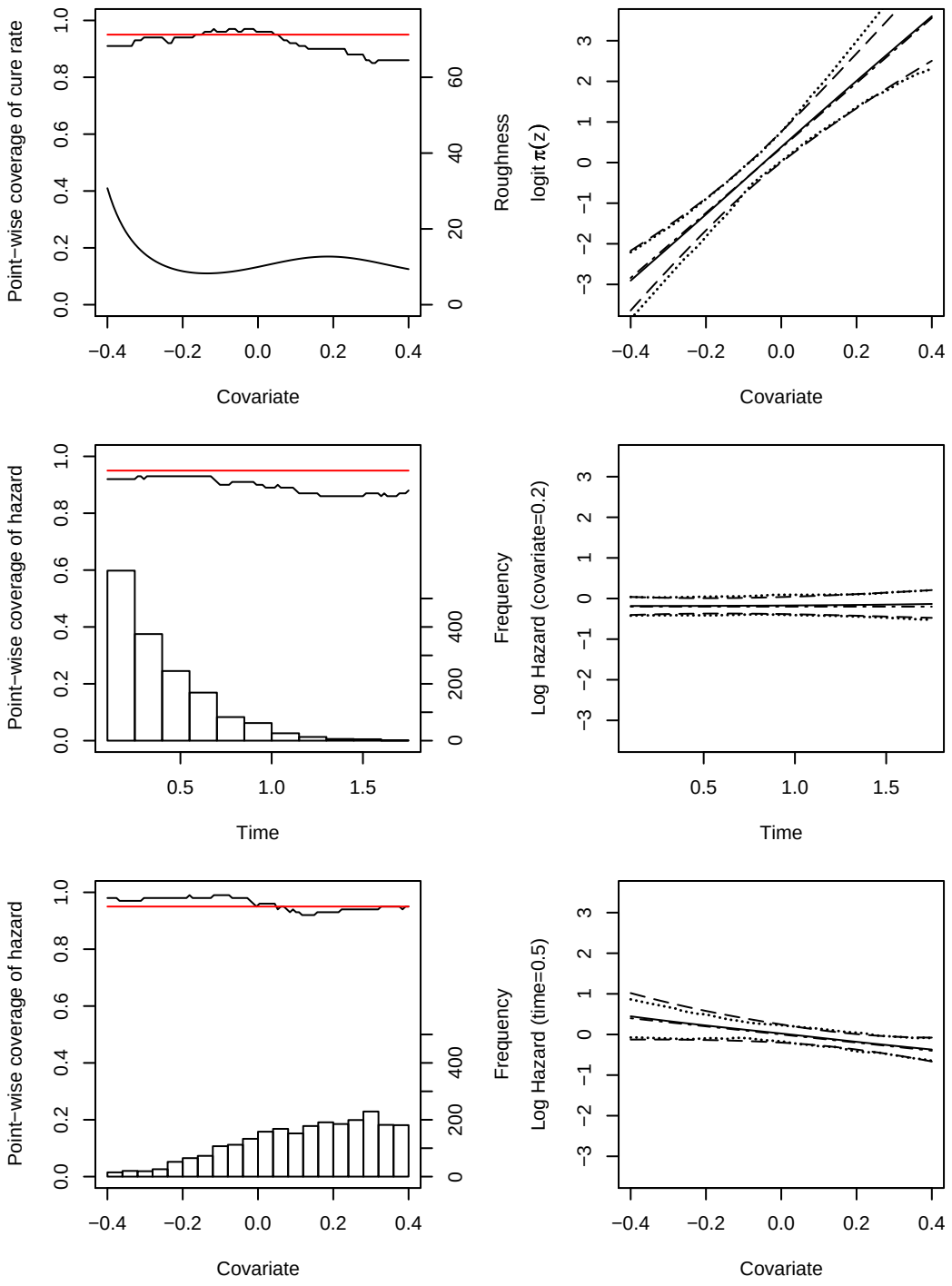

Fig. A.1. Simulation Results for Test Functions $\pi_4(z)$, $h_5(t,x)$ and $n = 400$. Left column: Point-wise coverages (top black lines). Superimposed are nominal coverage (red lines) and scaled $|logit(\pi''(z))|$ (bottom black line) or histogram of data. Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates.

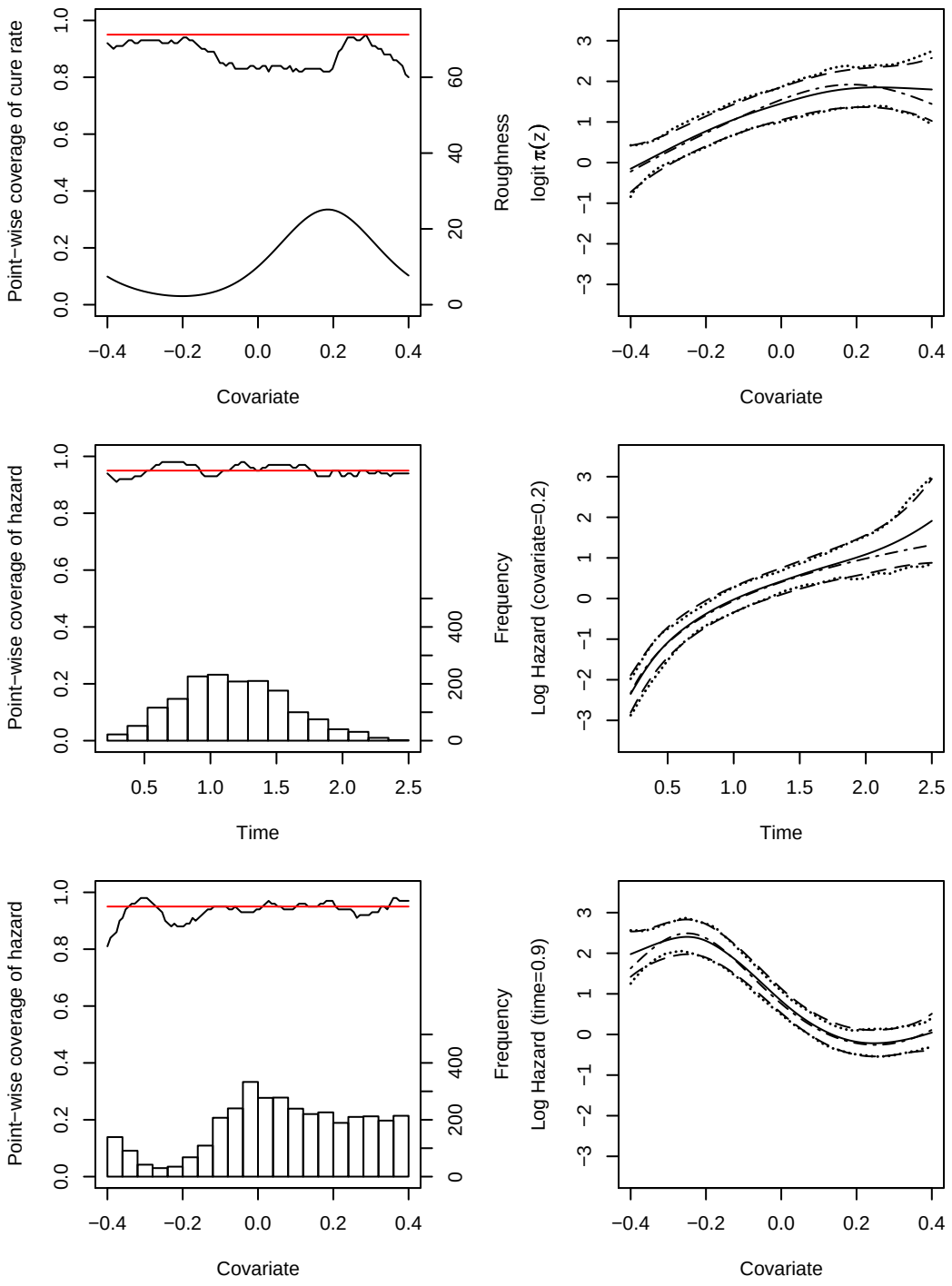

Fig. A.2. Simulation Results for Test Functions $\pi_1(z)$, $h_1(t,x)$ and $n = 400$. Left column: Point-wise coverages (top black lines). Superimposed are nominal coverage (red lines) and scaled $|logit(\pi''(z))|$ (bottom black line) or histogram of data. Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates.

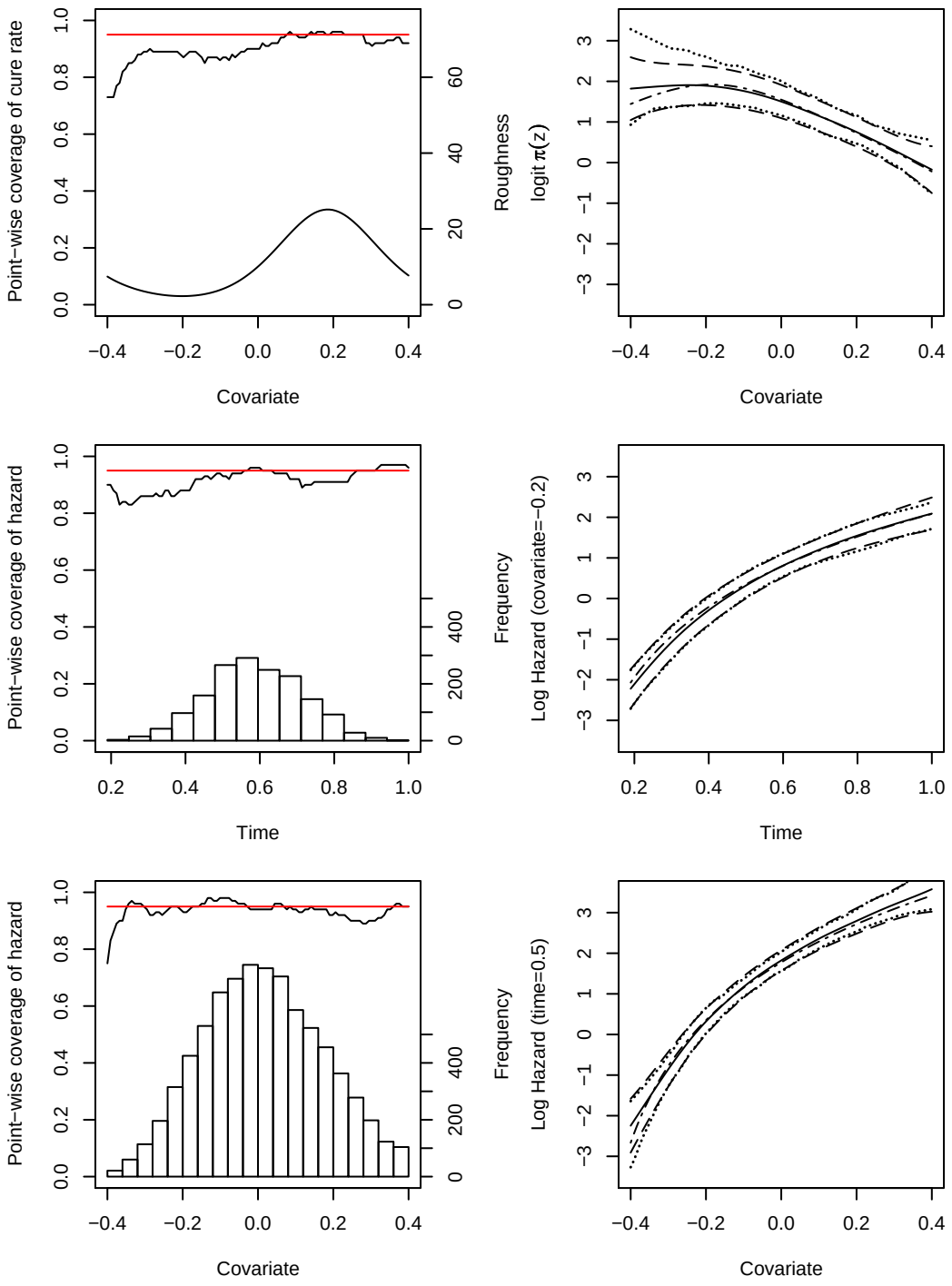

Fig. A.3. Simulation Results for Test Functions $\pi_1(z)$, $h_2(t,x)$ and $n = 400$. Left column: Point-wise coverages (top black lines). Superimposed are nominal coverage (red lines) and scaled $|logit(\pi''(z))|$ (bottom black line) or histogram of data. Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates.

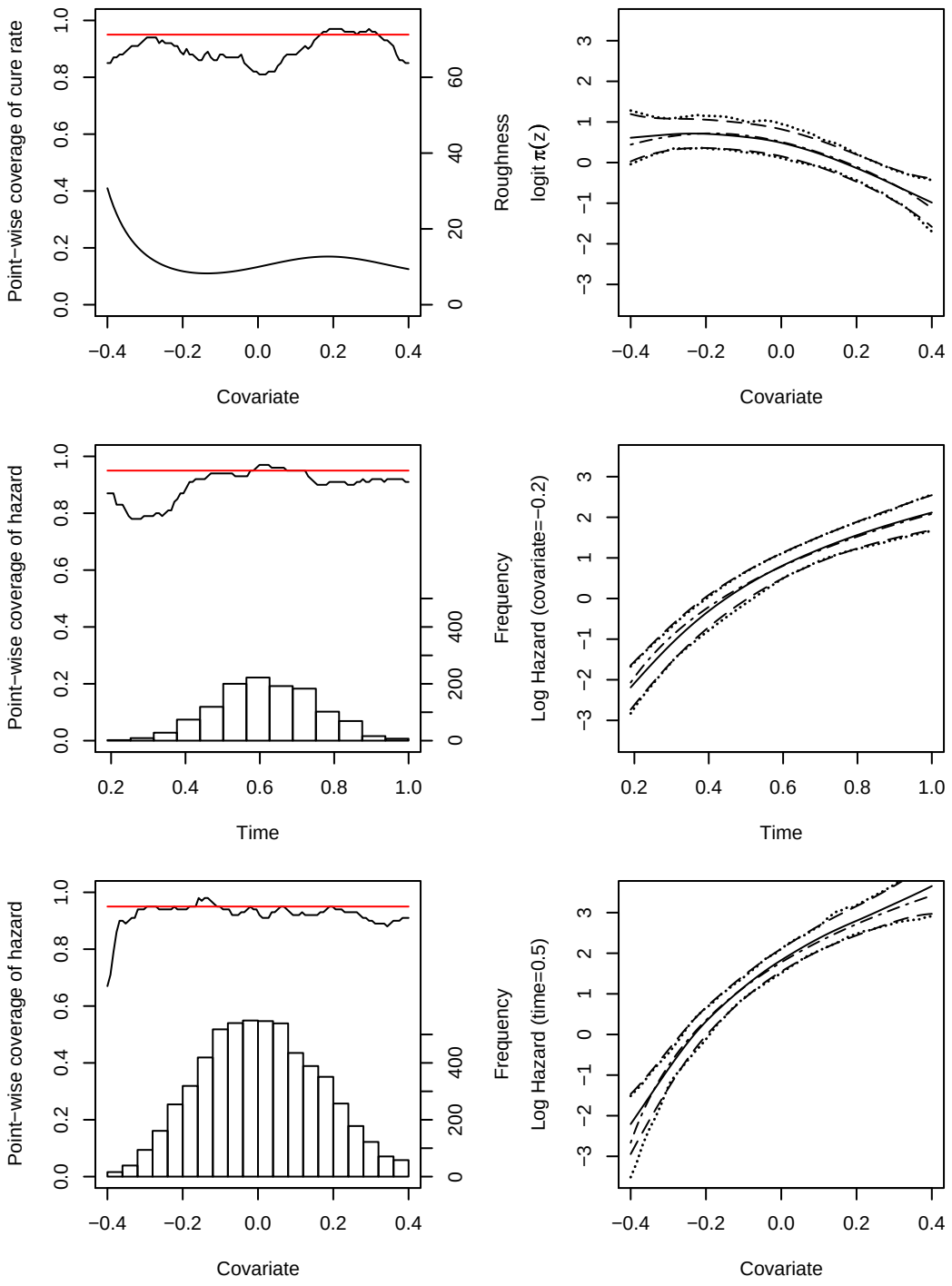

Fig. A.4. Simulation Results for Test Functions $\pi_2(z)$, $h_2(t,x)$ and $n = 400$. Left column: Point-wise coverages (top black lines). Superimposed are nominal coverage (red lines) and scaled $|logit(\pi''(z))|$ (bottom black line) or histogram of data. Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates.

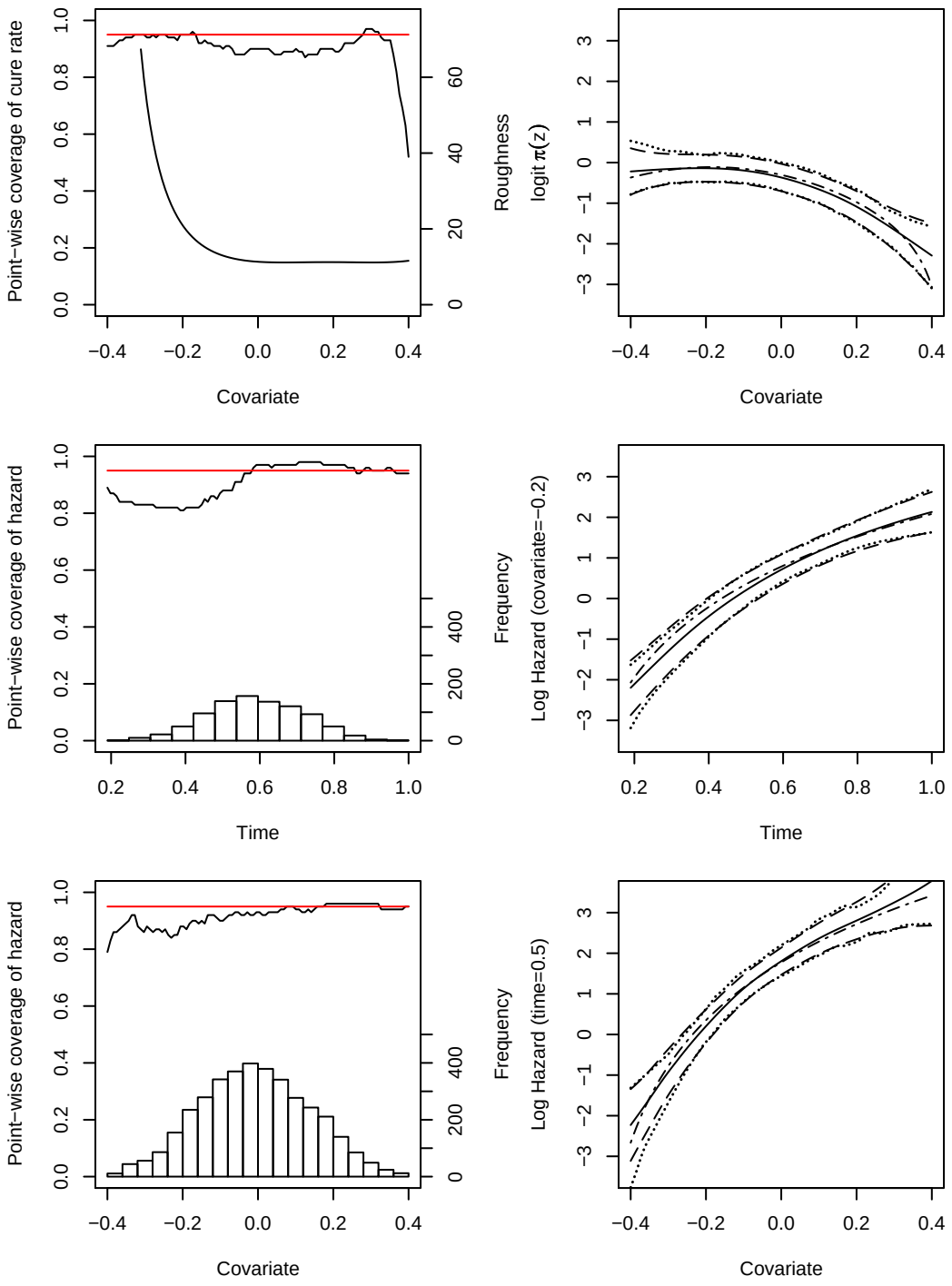

Fig. A.5. Simulation Results for Test Functions $\pi_3(z)$, $h_2(t,x)$ and $n = 400$. Left column: Point-wise coverages (top black lines). Superimposed are nominal coverage (red lines) and scaled $|logit(\pi''(z))|$ (bottom black line) or histogram of data. Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates.

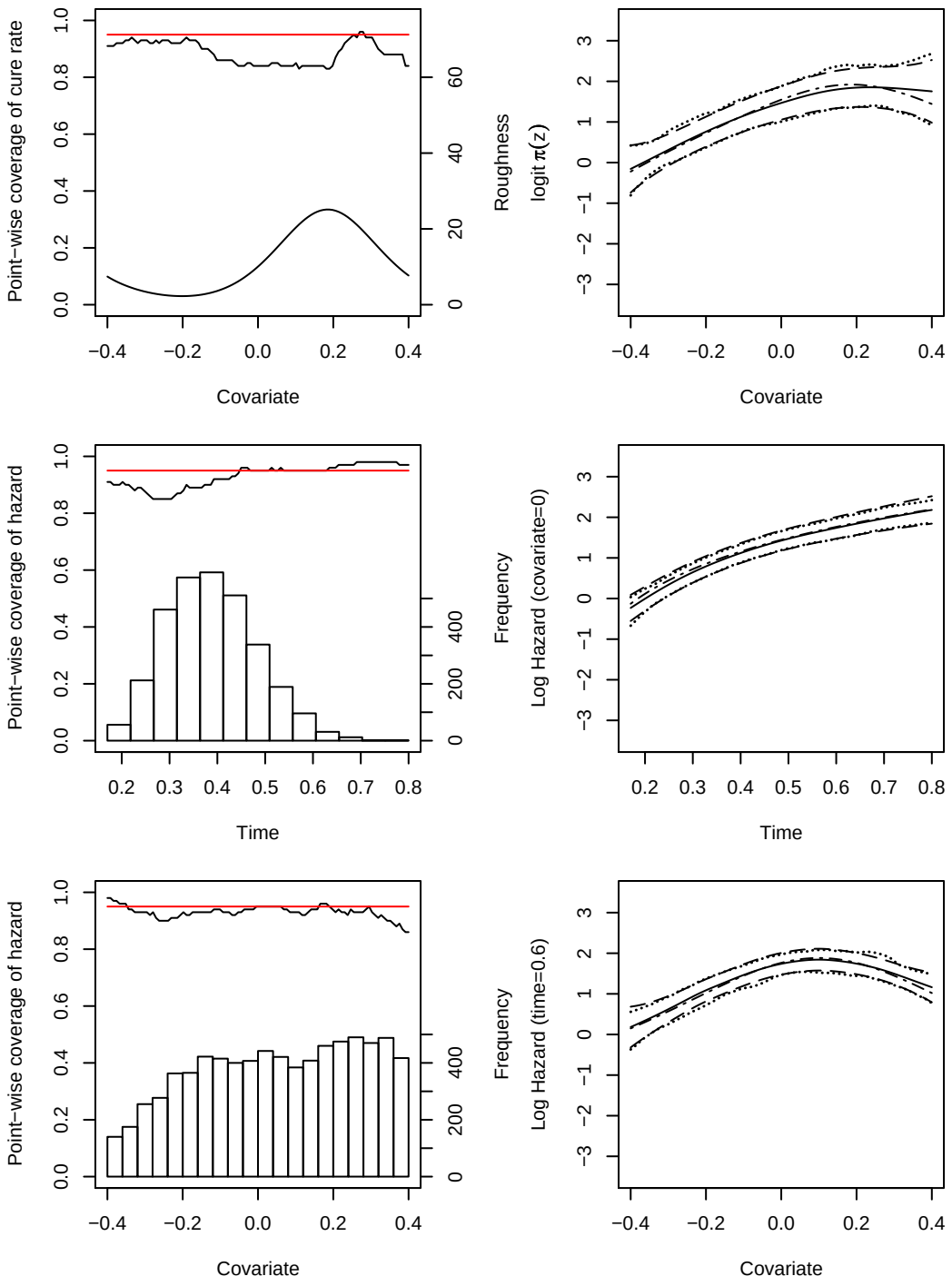

Fig. A.6. Simulation Results for Test Functions $\pi_1(z)$, $h_3(t,x)$ and $n = 400$. Left column: Point-wise coverages (top black lines). Superimposed are nominal coverage (red lines) and scaled $|logit(\pi''(z))|$ (bottom black line) or histogram of data. Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates.

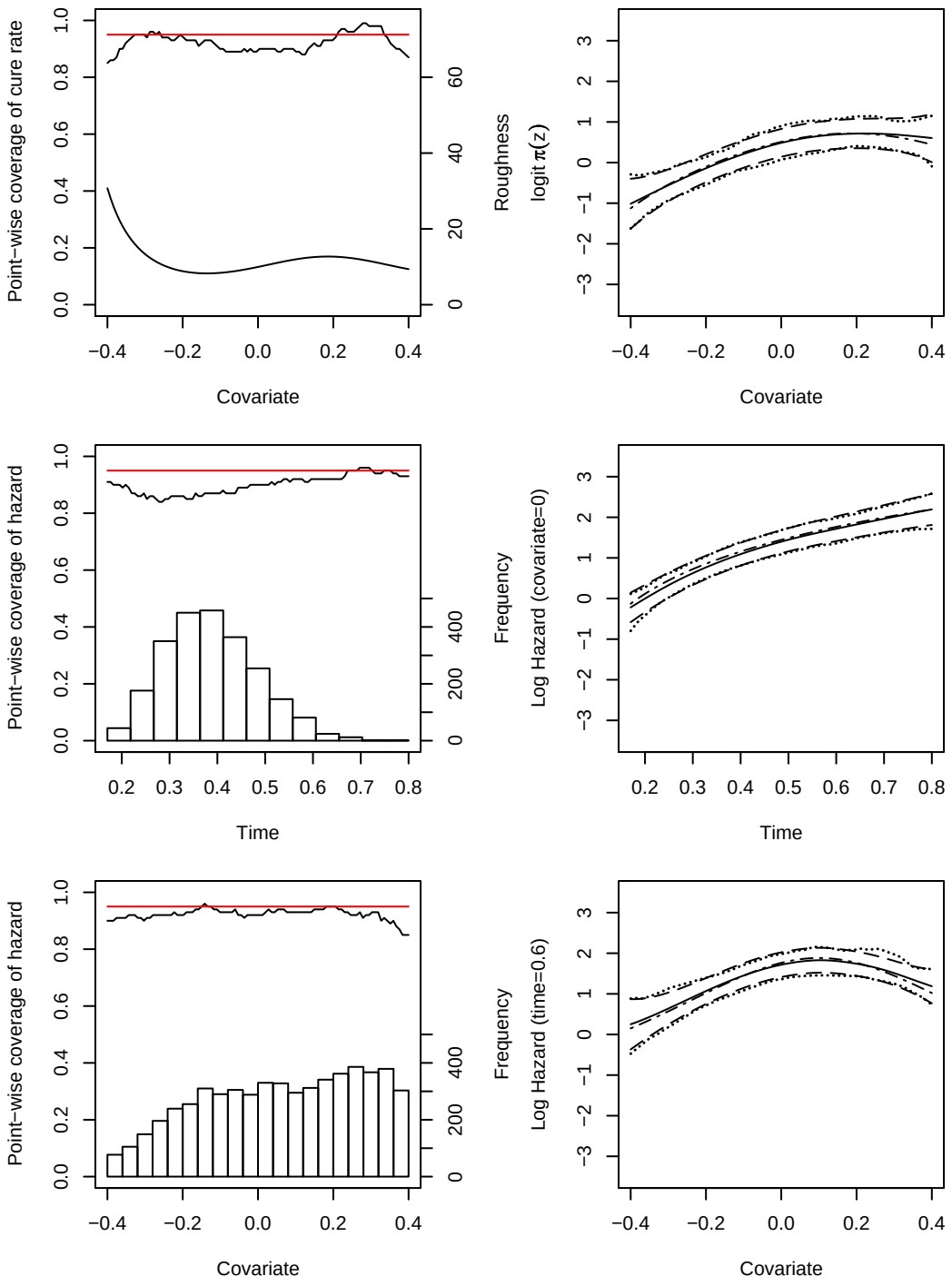

Fig. A.7. Simulation Results for Test Functions $\pi_2(z)$, $h_3(t,x)$ and $n = 400$. Left column: Point-wise coverages (top black lines). Superimposed are nominal coverage (red lines) and scaled $|logit(\pi''(z))|$ (bottom black line) or histogram of data. Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates.

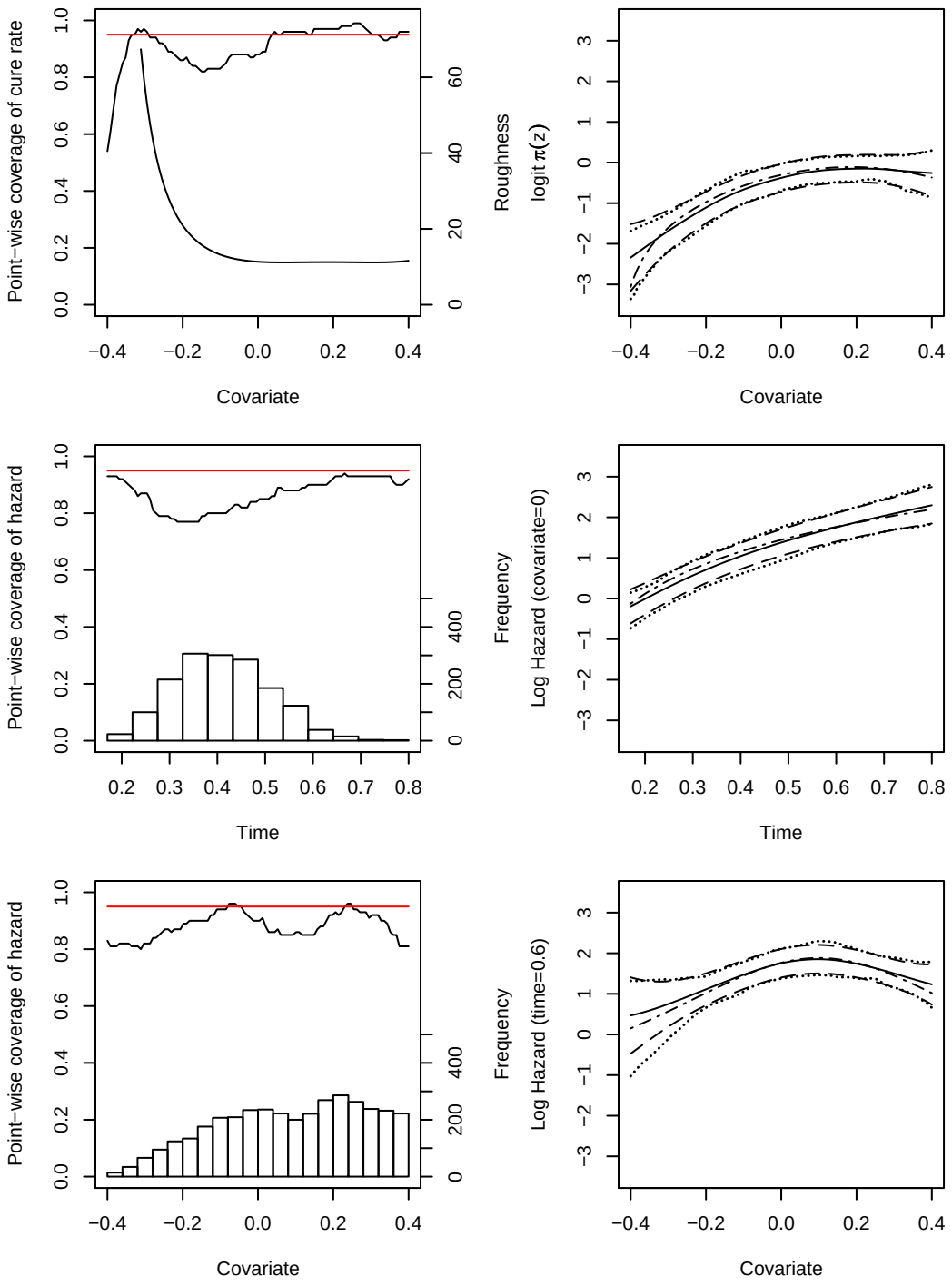

Fig. A.8. Simulation Results for Test Functions $\pi_3(z)$, $h_3(t,x)$ and $n = 400$. Left column: Point-wise coverages (top black lines). Superimposed are nominal coverage (red lines) and scaled $|logit(\pi''(z))|$ (bottom black line) or histogram of data. Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates.

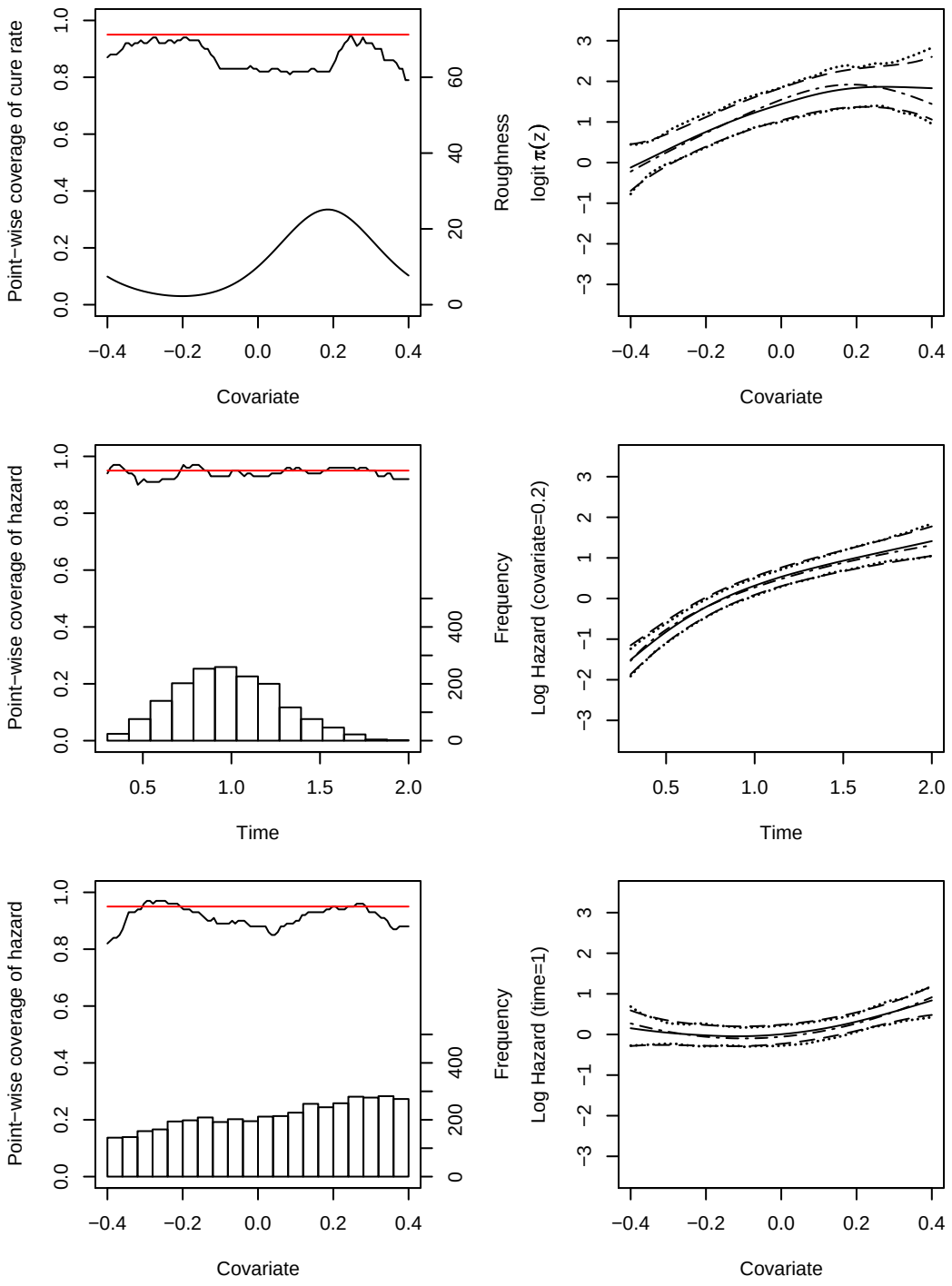

Fig. A.9. Simulation Results for Test Functions $\pi_1(z)$, $h_4(t,x)$ and $n = 400$. Left column: Point-wise coverages (top black lines). Superimposed are nominal coverage (red lines) and scaled $|logit(\pi''(z))|$ (bottom black line) or histogram of data. Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates.

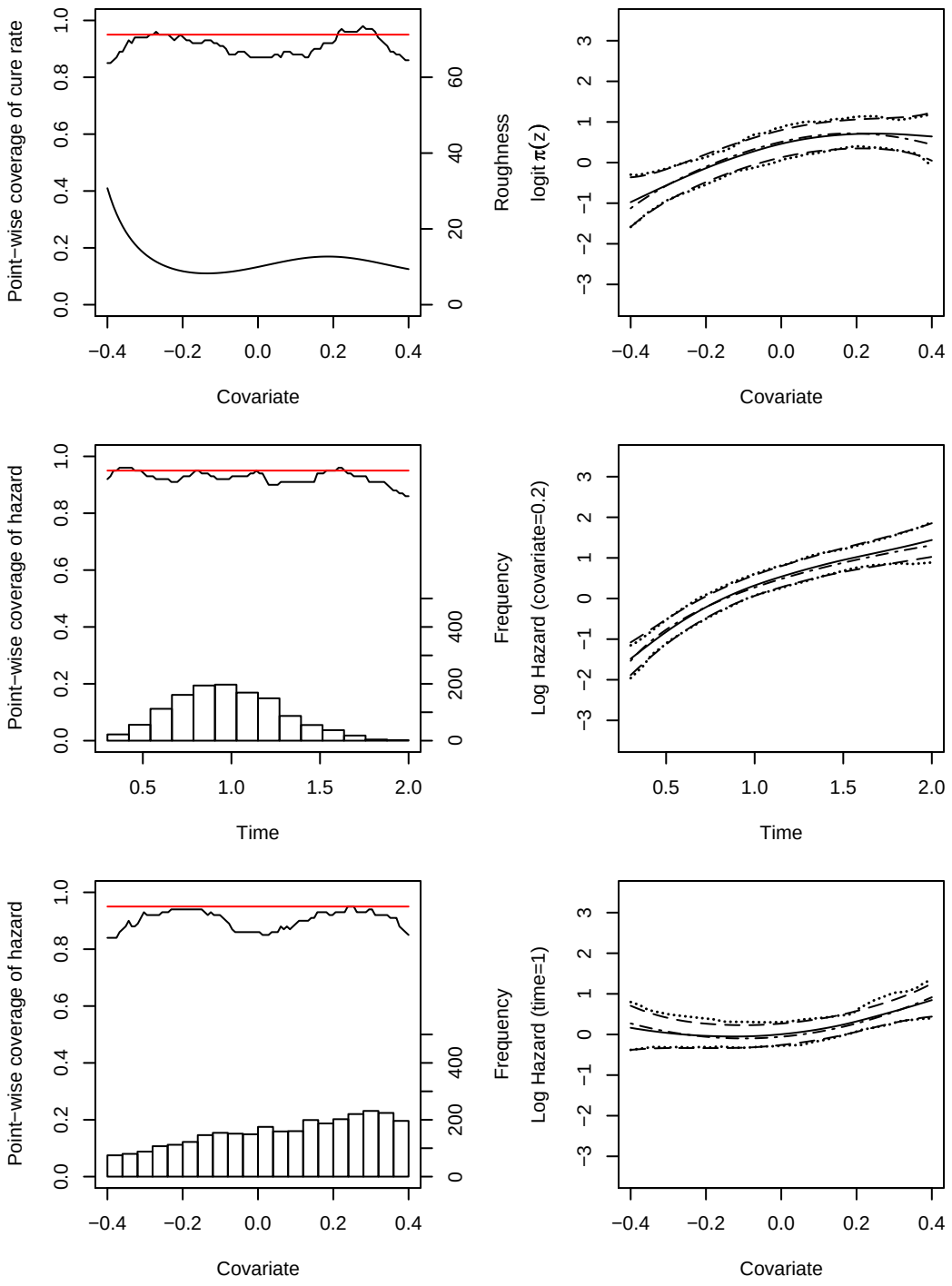

Fig. A.10. Simulation Results for Test Functions $\pi_2(z)$, $h_4(t,x)$ and $n = 400$. Left column: Point-wise coverages (top black lines). Superimposed are nominal coverage (red lines) and scaled $|logit(\pi''(z))|$ (bottom black line) or histogram of data. Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates.

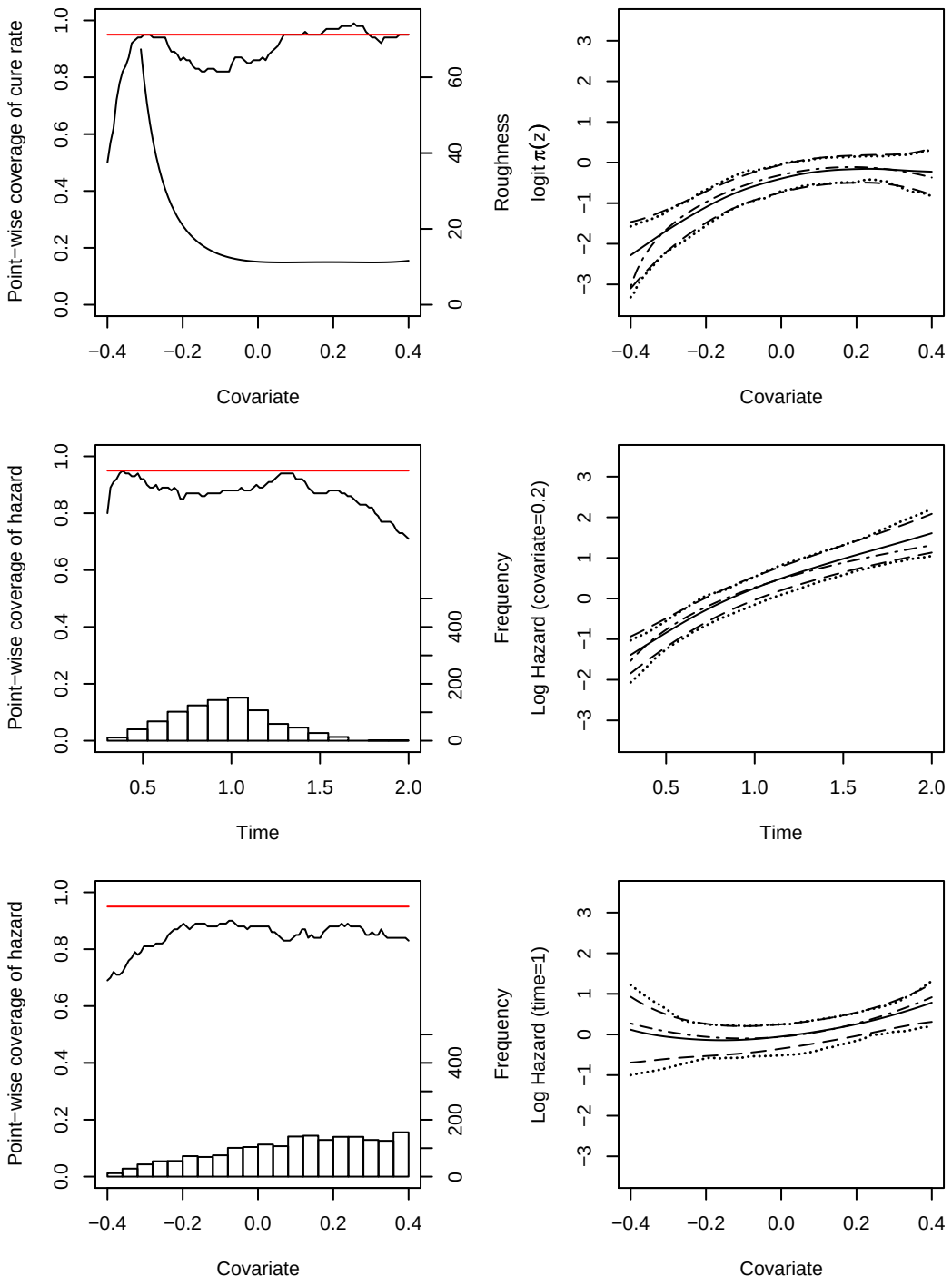

Fig. A.11. Simulation Results for Test Functions $\pi_3(z)$, $h_4(t,x)$ and $n = 400$. Left column: Point-wise coverages (top black lines). Superimposed are nominal coverage (red lines) and scaled $|logit(\pi''(z))|$ (bottom black line) or histogram of data. Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates.

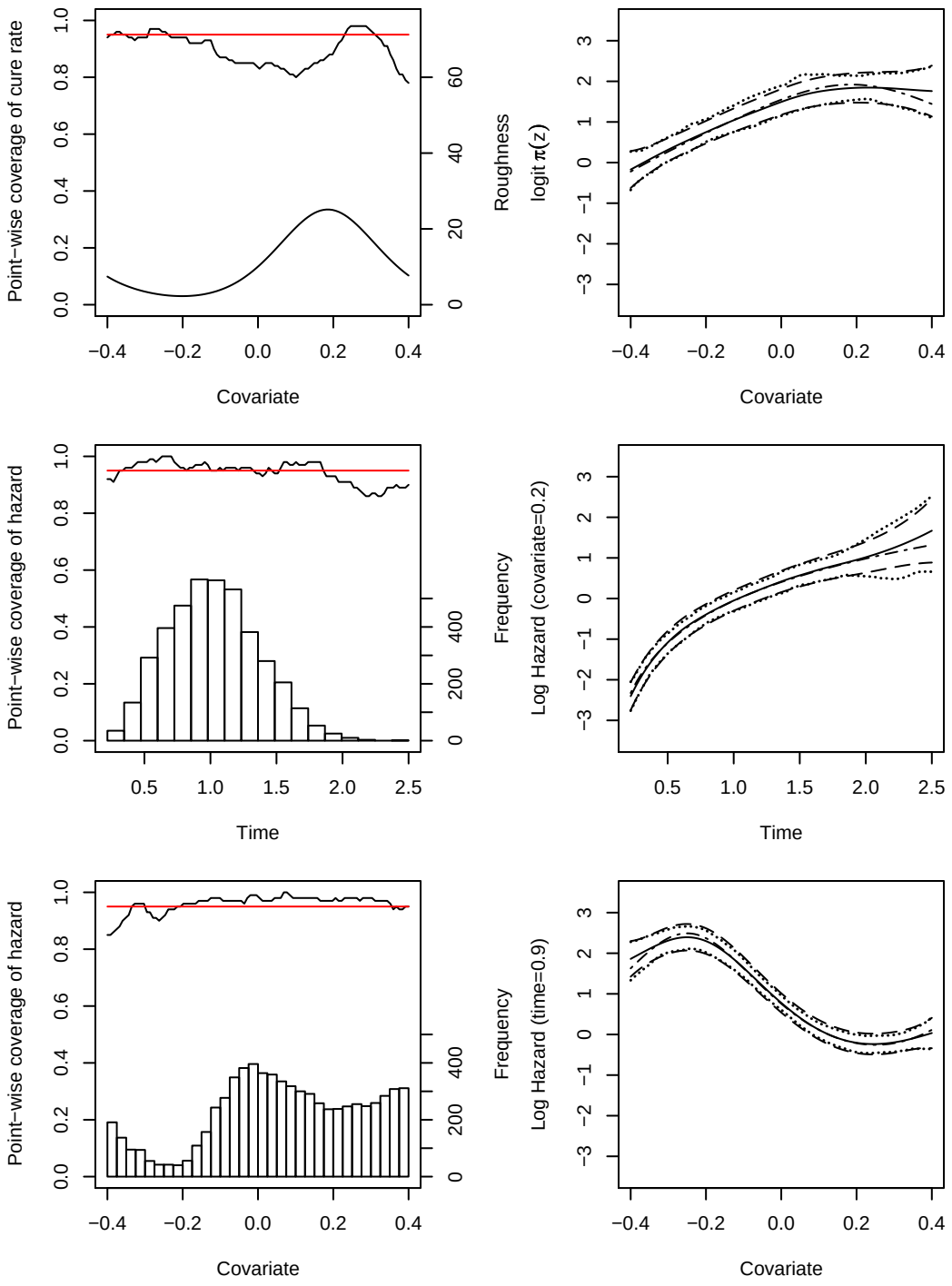

Fig. A.12. Simulation Results for Test Functions $\pi_1(z)$, $h_1(t,x)$ and $n = 800$. Left column: Point-wise coverages (top black lines). Superimposed are nominal coverage (red lines) and scaled $|logit(\pi''(z))|$ (bottom black line) or histogram of data. Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates.

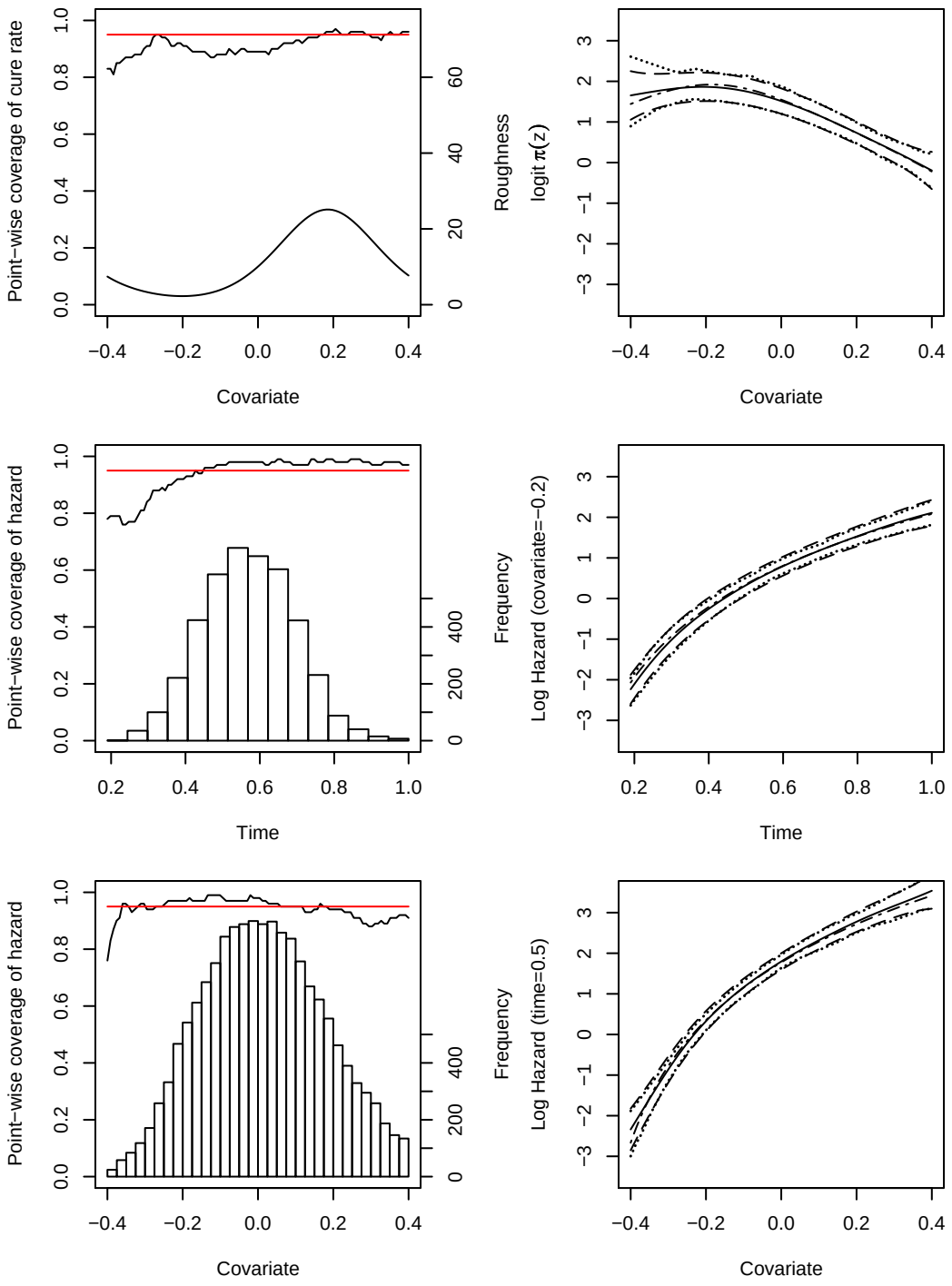

Fig. A.13. Simulation Results for Test Functions $\pi_1(z)$, $h_2(t,x)$ and $n = 800$. Left column: Point-wise coverages (top black lines). Superimposed are nominal coverage (red lines) and scaled $|logit(\pi''(z))|$ (bottom black line) or histogram of data. Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates.

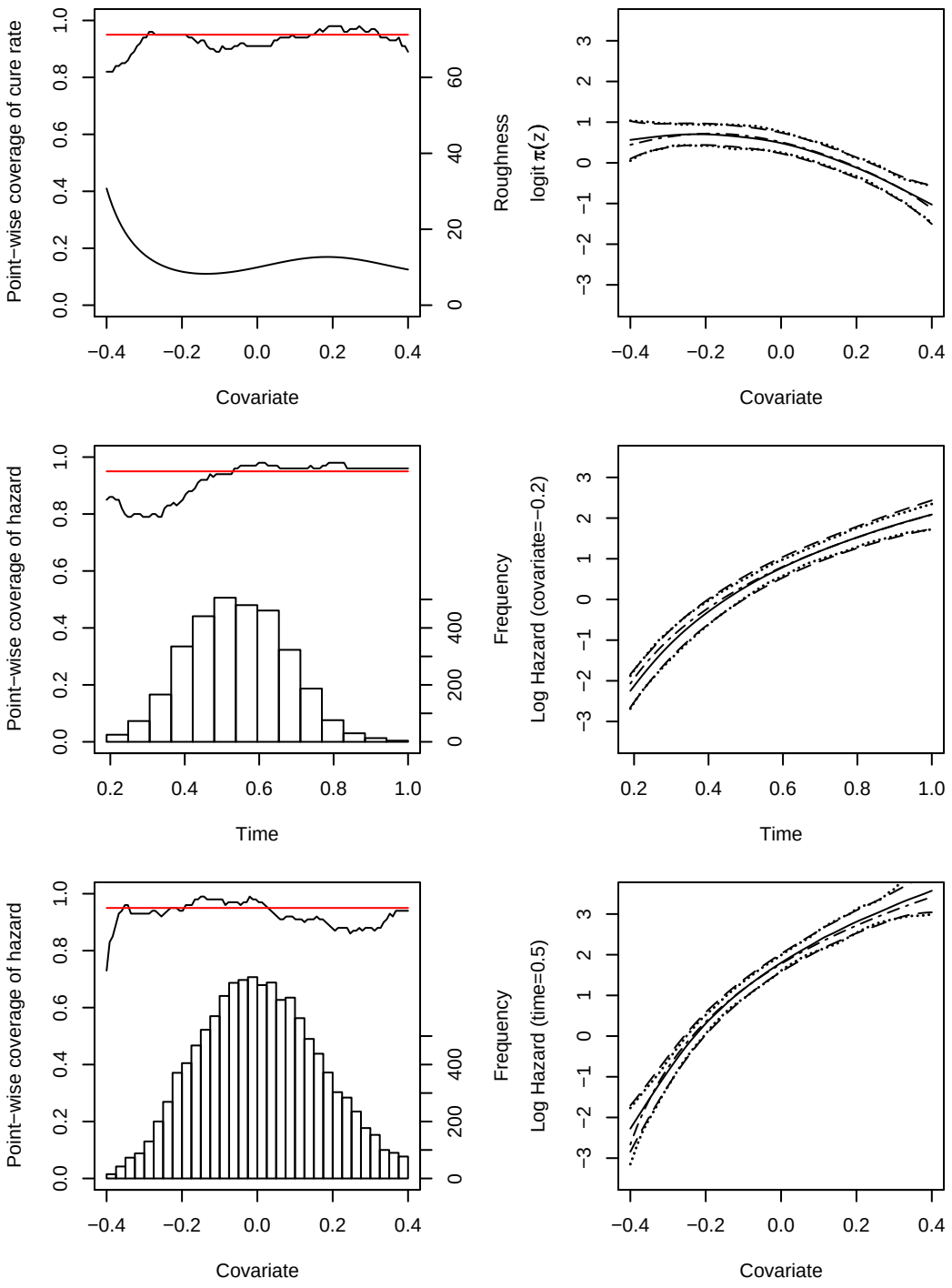

Fig. A.14. Simulation Results for Test Functions $\pi_2(z)$, $h_2(t,x)$ and $n = 800$. Left column: Point-wise coverages (top black lines). Superimposed are nominal coverage (red lines) and scaled $|logit(\pi''(z))|$ (bottom black line) or histogram of data. Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates.

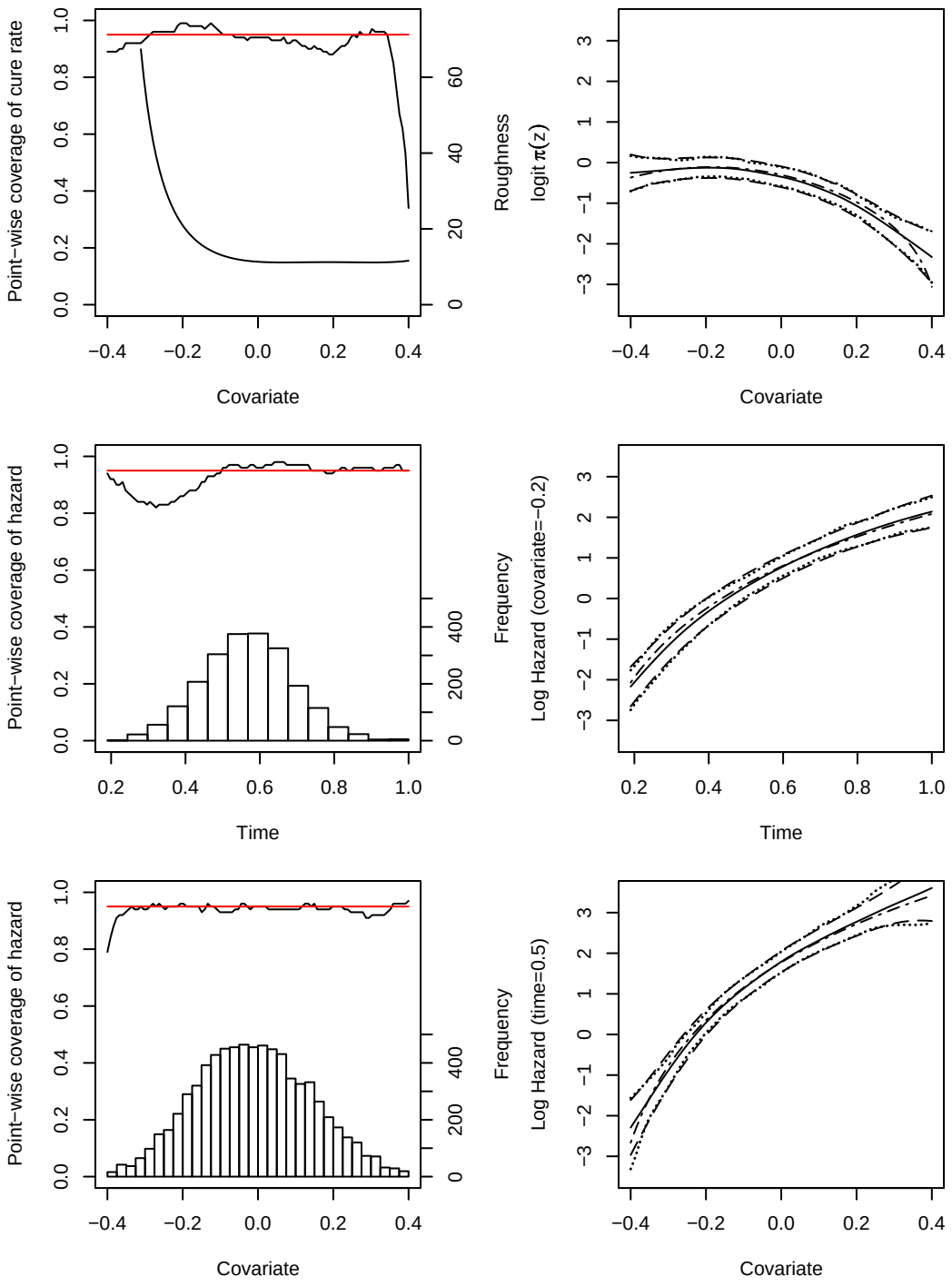

Fig. A.15. Simulation Results for Test Functions $\pi_3(z)$, $h_2(t,x)$ and $n = 800$. Left column: Point-wise coverages (top black lines). Superimposed are nominal coverage (red lines) and scaled $|logit(\pi''(z))|$ (bottom black line) or histogram of data. Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates.

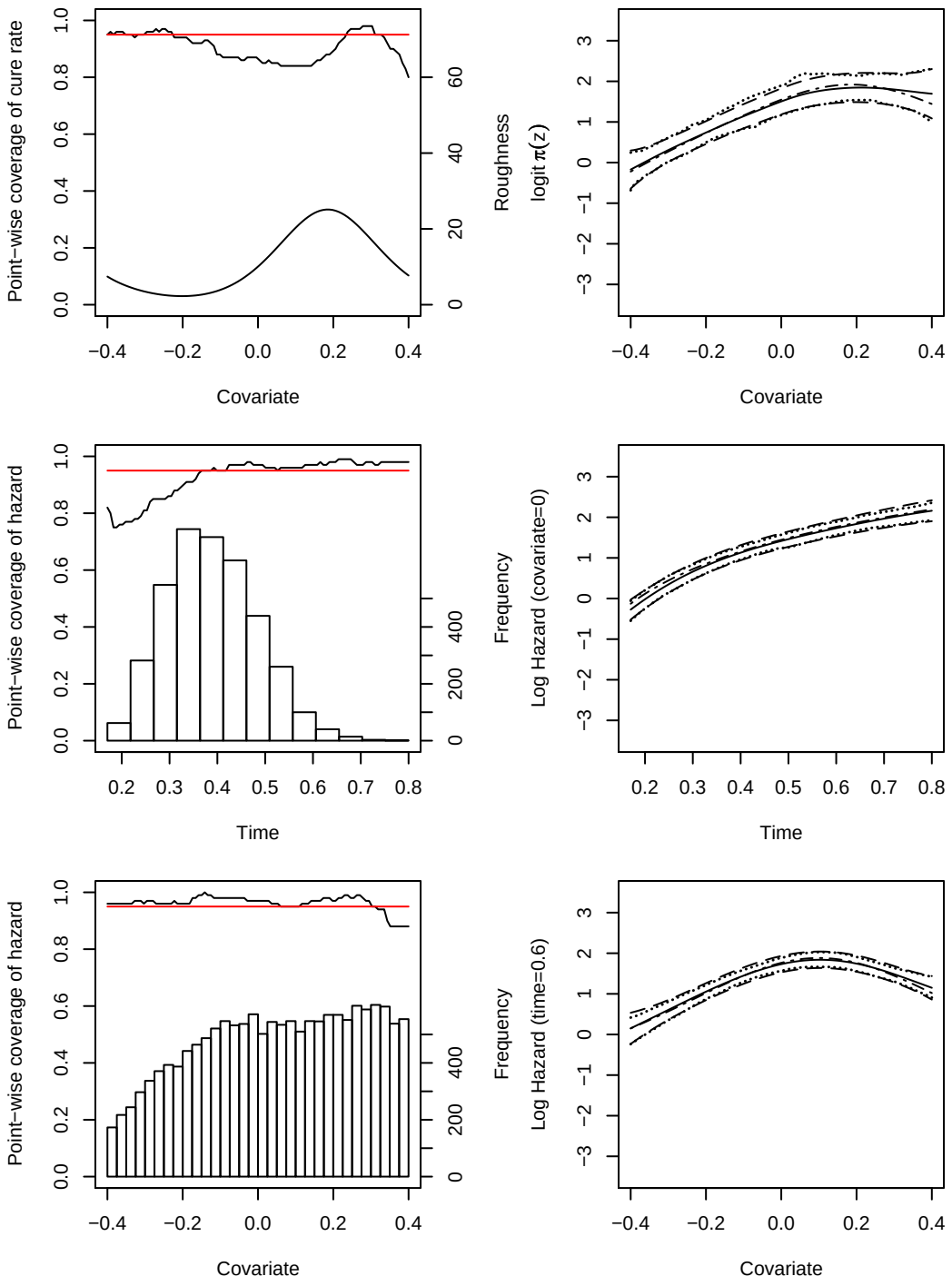

Fig. A.16. Simulation Results for Test Functions $\pi_1(z)$, $h_3(t,x)$ and $n = 800$. Left column: Point-wise coverages (top black lines). Superimposed are nominal coverage (red lines) and scaled $|logit(\pi''(z))|$ (bottom black line) or histogram of data. Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates.

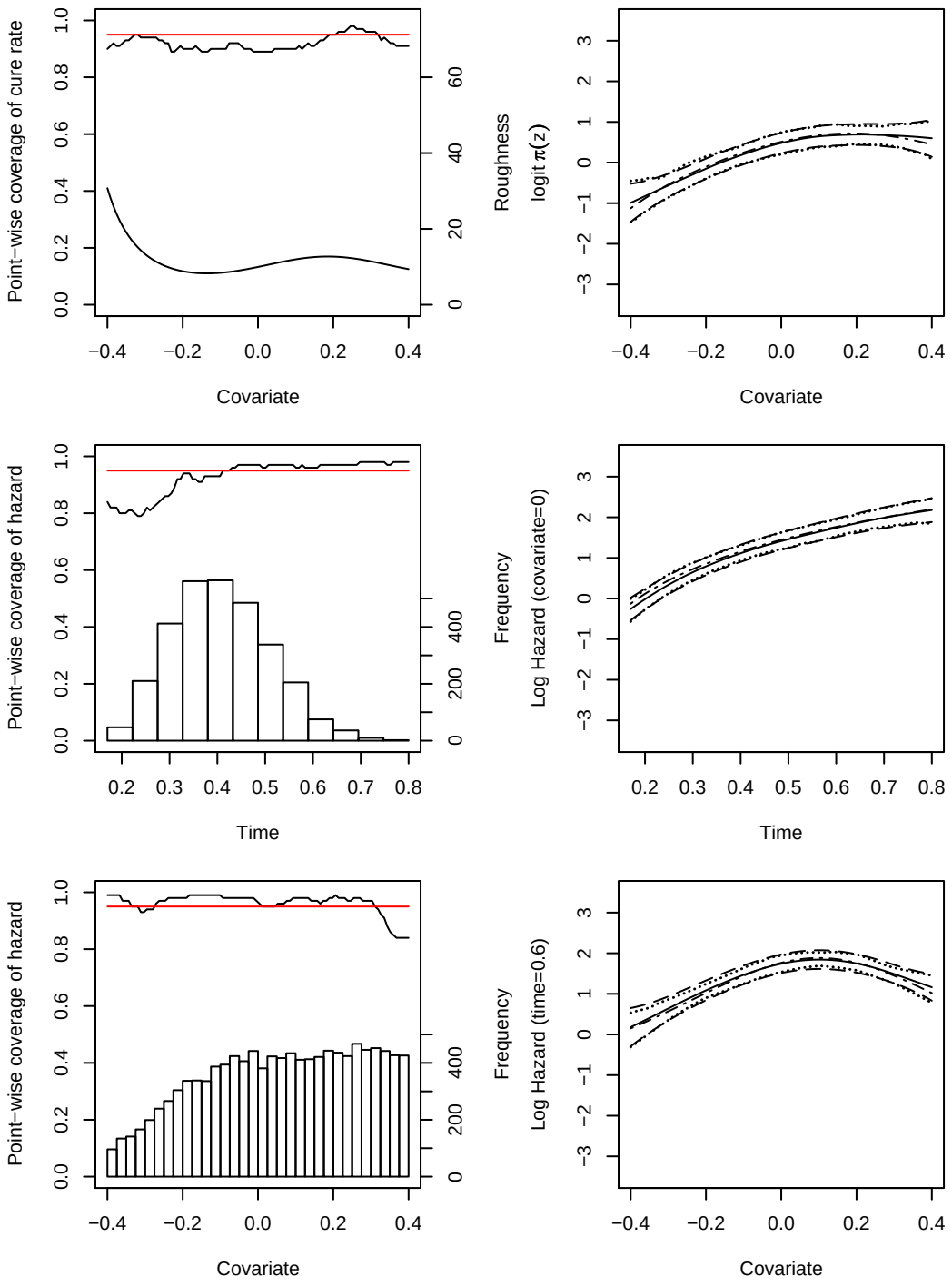

Fig. A.17. Simulation Results for Test Functions $\pi_2(z)$, $h_3(t,x)$ and $n = 800$. Left column: Point-wise coverages (top black lines). Superimposed are nominal coverage (red lines) and scaled $|logit(\pi''(z))|$ (bottom black line) or histogram of data. Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates.

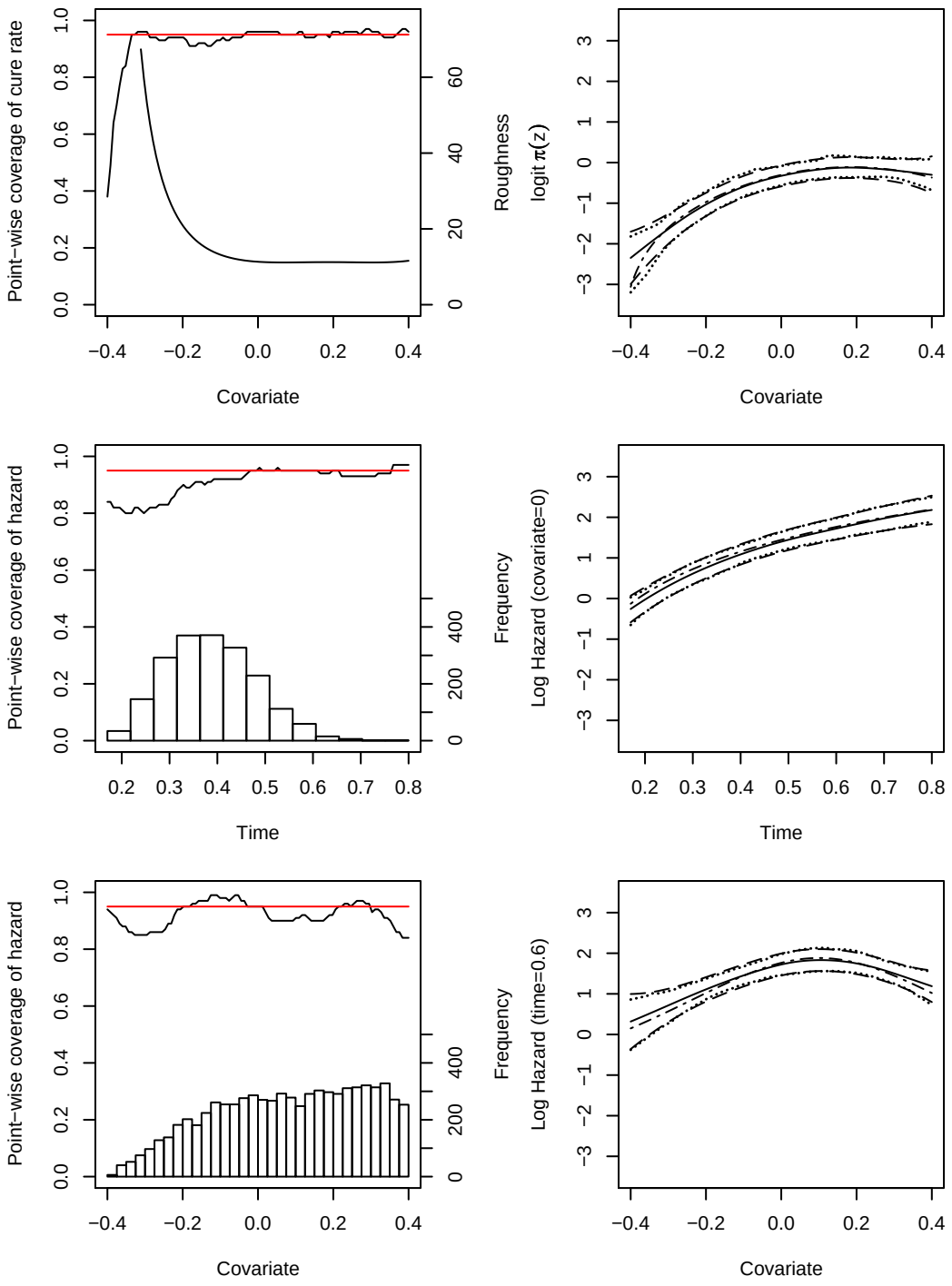

Fig. A.18. Simulation Results for Test Functions $\pi_3(z)$, $h_3(t,x)$ and $n = 800$. Left column: Point-wise coverages (top black lines). Superimposed are nominal coverage (red lines) and scaled $|logit(\pi''(z))|$ (bottom black line) or histogram of data. Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates.

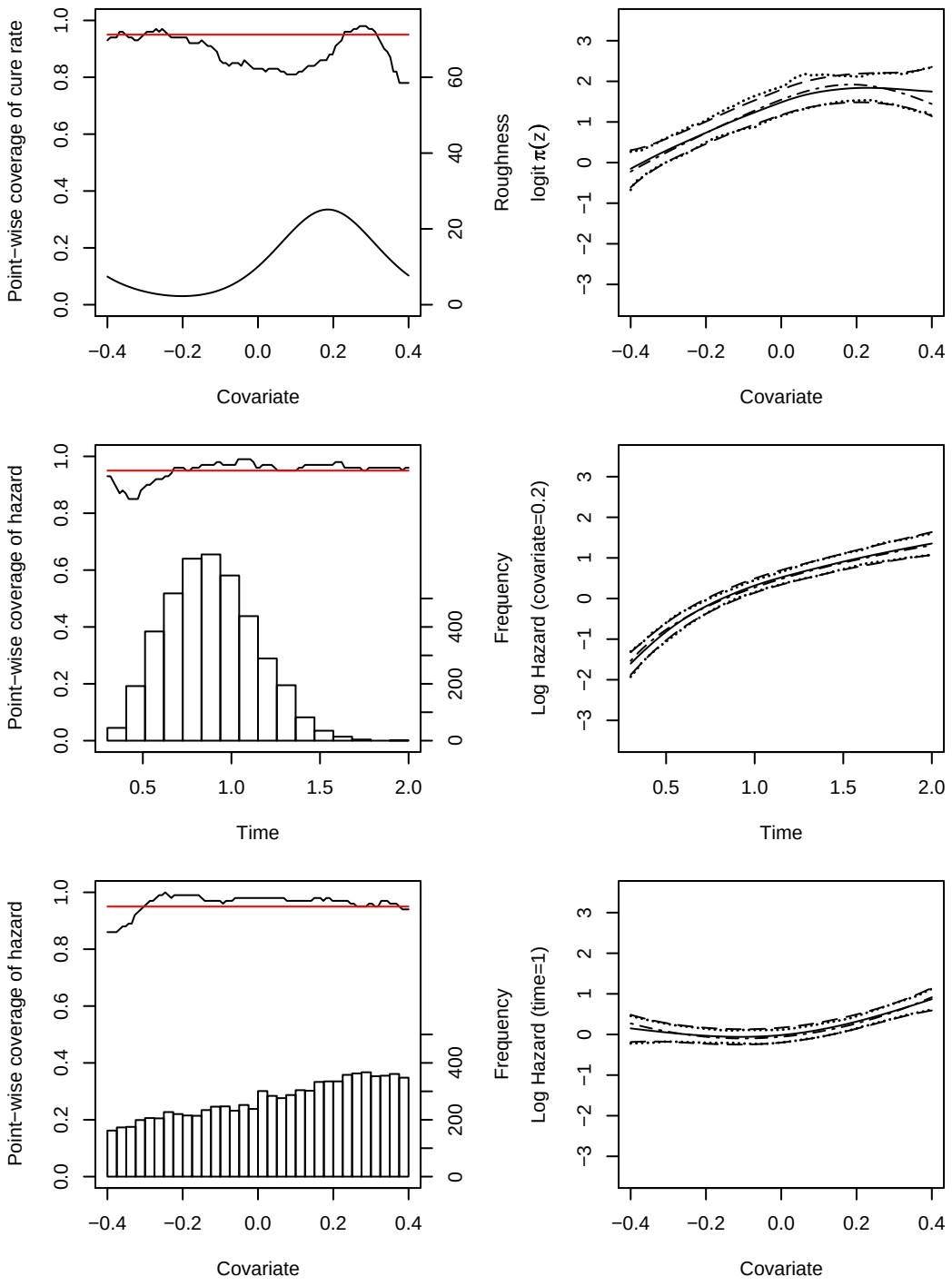

Fig. A.19. Simulation Results for Test Functions $\pi_1(z)$, $h_4(t,x)$ and $n = 800$. Left column: Point-wise coverages (top black lines). Superimposed are nominal coverage (red lines) and scaled $|logit(\pi''(z))|$ (bottom black line) or histogram of data. Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates.

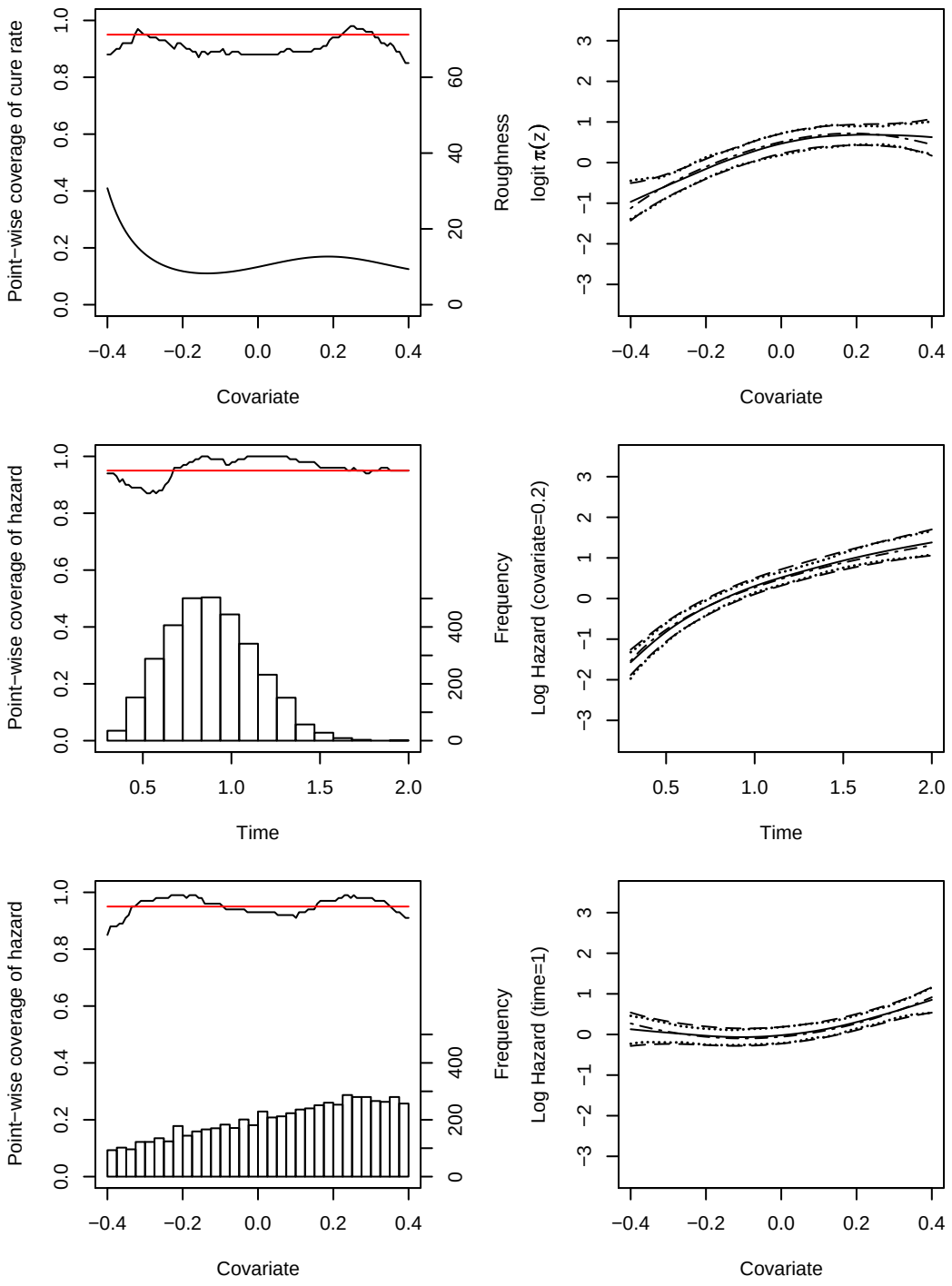

Fig. A.20. Simulation Results for Test Functions $\pi_2(z)$, $h_4(t,x)$ and $n = 800$. Left column: Point-wise coverages (top black lines). Superimposed are nominal coverage (red lines) and scaled $|logit(\pi''(z))|$ (bottom black line) or histogram of data. Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates.

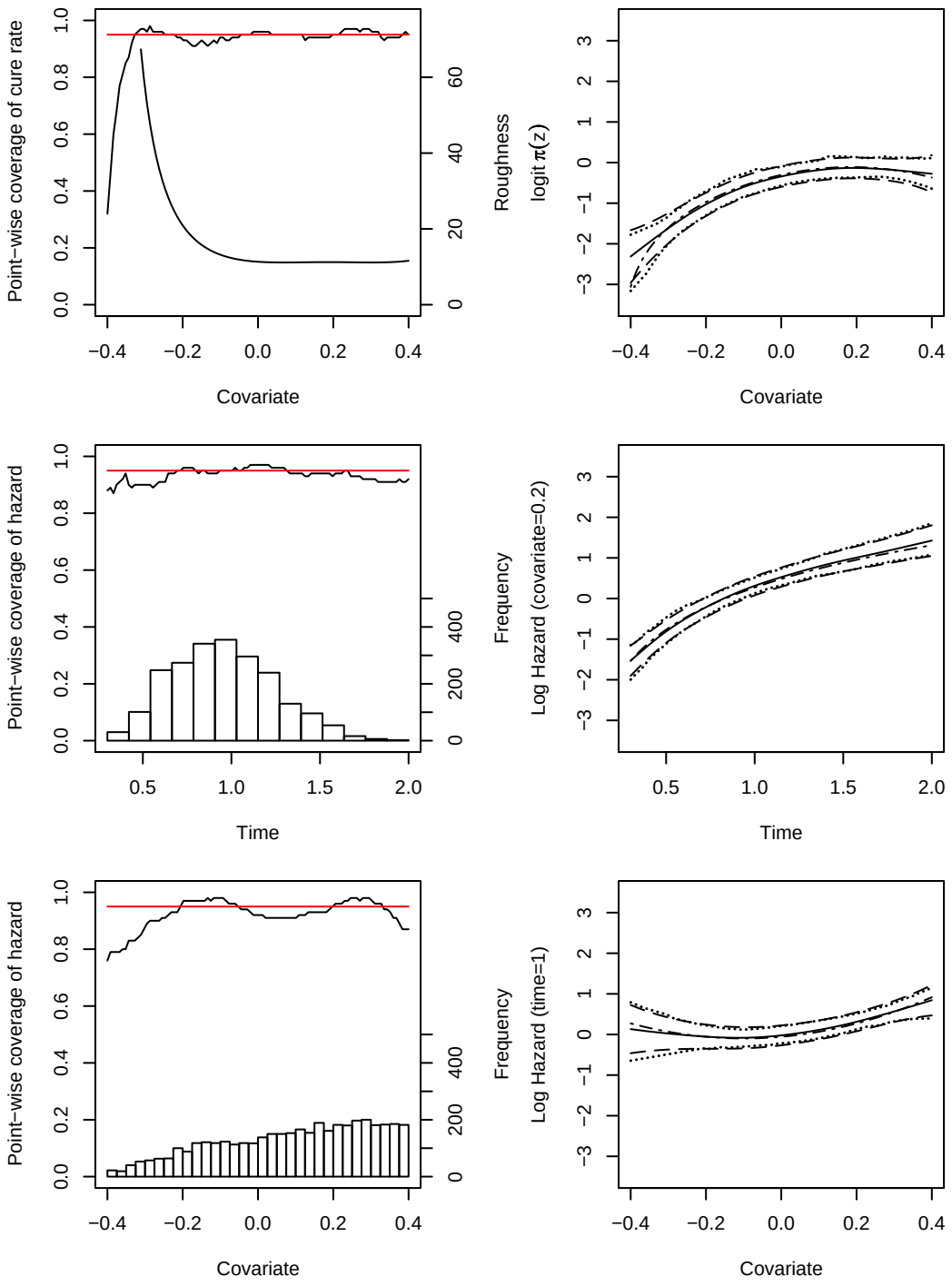

Fig. A.21. Simulation Results for Test Functions $\pi_3(z)$, $h_4(t,x)$ and $n = 800$. Left column: Point-wise coverages (top black lines). Superimposed are nominal coverage (red lines) and scaled $|logit(\pi''(z))|$ (bottom black line) or histogram of data. Right column: True functions (dash-dotted) and their estimates, including averages of point-wise function estimates (solid), averages of point-wise 95% CIs (dashed) and empirical 2.5 and 97.5 percentiles of point-wise function estimates (dotted), all based on 100 data replicates.