Statistica Sinica **8**(1998), 317-335

STRUCTURAL MULTIVARIATE FUNCTION ESTIMATION: SOME AUTOMATIC DENSITY AND HAZARD ESTIMATES

Chong Gu

Purdue University

Abstract: Structures such as independence of random variables in probability densities and hazard proportionality in covariate dependent hazard functions have important interpretations in statistical analysis. Such structures can be characterized by term eliminations from an analysis of variance (ANOVA) decomposition in log density or log hazard. Nonparametric estimation of these functions with an ANOVA decomposition built in can be achieved by using tensor product splines in a penalized likelihood approach. In this article, a feasible algorithm with automatic multiple smoothing parameters is described to implement this approach, and examples are presented to illustrate some applications of the technique. For density estimation, a novel feature is the possibility of assessing/enforcing independence when data are truncated to a non rectangular domain. For hazard estimation, models more general than but reducible to proportional hazard models are available, and model terms are estimated simultaneously via penalized full likelihood.

Key words and phrases: Analysis of variance decomposition, penalized likelihood method, performance-oriented iteration, smoothing parameter, tensor product spline.

1. Introduction

Data and models are two sources of information in a statistical analysis. Data carry noise but are "unbiased", whereas models, or constraints, help to reduce noise but are responsible for "biases". Parametric restrictive models and constraint-free nonparametric analyses (e.g., the empirical distribution for densities and the Kaplan-Meier estimator for hazards) represent two extremes on the spectrum of bias-variance tradeoff. Smooth function models with soft constraints come in between the two extremes. Among the many smoothing methods available, the penalized likelihood method pioneered by Good and Gaskins (1971) allows convenient structural model construction, and hence is rather handy for handling multivariate problems.

Let η be a function of interest. Smooth models for η can be specified via $M_\rho = \{\eta : J(\eta) \leq \rho\}$, where $J(\eta)$ is a quadratic roughness functional with a low dimensional null space $J_\perp$. An example of $J(\eta)$ on an interval, say $[0, 1]$, is $\int \ddot{\eta}^2 dx$. When $\rho = 0$, $M_0 = J_\perp$ defines a parametric model. As ρ increases, M_ρ

allows more and more flexible fits. To fit the model in M_ρ, one usually resorts to the maximum likelihood method. The estimate in $\{\eta : J(\eta) \le \rho\}$ usually falls on the sphere $\{\eta : J(\eta) = \rho\}$, and Lagrange's method turns the problem into a penalized likelihood problem

$$\min\ L(\eta) + (\lambda/2)J(\eta), \tag{1.1}$$

where $L(\eta)$ is usually the minus log likelihood of the data. The Lagrange multiplier λ is called the smoothing parameter, which controls the tradeoff between the goodness-of-fit and the smoothness of η.

A few generic examples of penalized likelihood estimation follow.

Example 1.1. *Response Data Regression.* Assume $Y|X \sim \exp\{(y\eta(x) - b(\eta(x)))/a(\phi) + c(y,\phi)\}$, an exponential family density with a modeling parameter η and a possibly unknown nuisance parameter ϕ. Observing independent data (X_i, Y_i), $i = 1, \ldots, n$, the method estimates η via minimizing

$$-\frac{1}{n}\sum_{i=1}^{n}\{Y_i\eta(X_i) - b(\eta(X_i))\} + \frac{\lambda}{2}J(\eta). \tag{1.2}$$

Example 1.2. *Density Estimation.* Observing i.i.d. samples X_i, $i = 1, \ldots, n$, from a probability density $f(x)$ supported on a finite domain $\mathcal{X}$, the method estimates f by $e^\eta / \int_{\mathcal{X}} e^\eta dx$, where η minimizes

$$-\frac{1}{n}\sum_{i=1}^{n}\Big\{\eta(X_i) - \log\int_{\mathcal{X}} e^\eta dx\Big\} + \frac{\lambda}{2}J(\eta). \tag{1.3}$$

A side condition, say $\int_{\mathcal{X}} \eta dx = 0$, shall be imposed on η for a one-to-one transform $f \leftrightarrow e^\eta / \int_{\mathcal{X}} e^\eta dx$.

Example 1.3. *Hazard Estimation.* Let T be the life time of an item with a survival function $S(t,u) = P(T > t|u)$, possibly dependent on a covariate u, and a hazard function $e^{\eta(t,u)} = -d\log S(t,u)/dt$. Let Z be the truncation time and C be the censoring time, independent of T and of each other. Observing $(Z_i, X_i, \delta_i, U_i)$, $i = 1, \ldots, n$, where $X = \min(T, C)$, $\delta = I_{[T \le C]}$, and $Z < X$, the method estimates the log hazard η via minimizing

$$-\frac{1}{n}\sum_{i=1}^{n}\Big\{\delta_i\eta(X_i, U_i) - \int_{Z_i}^{X_i} e^{\eta(t,U_i)}dt\Big\} + \frac{\lambda}{2}J(\eta). \tag{1.4}$$

Normal data regression, an important special case of Example 1.1, is by far the most intensively studied in the literature; a nice synthesis can be found in Wahba (1990). The formulation (1.2) for regression with general exponential

family data was proposed by O'Sullivan, Yandell and Raynor (1986); see also Silverman (1978). The formulation (1.3) appearing in Gu and Qiu (1993) evolved from the work of Good and Gaskins (1971), Leonard (1978), and Silverman (1982). The formulation (1.4) of Gu (1996) is influenced by the work of O'Sullivan (1988a, b) and Zucker and Karr (1990), among others.

Note that there is no dimensional restriction on x in Examples 1.1 and 1.2 so in general the problem could be a multivariate one. Example 1.3 is by definition multivariate unless the covariate domain reduces to a singleton. Structures based on a certain ANOVA decomposition of multivariate functions often help to enhance the interpretability of the estimates, and selective term trimming in such a decomposition may also help to partly ease the curse of dimensionality in estimation. As a simple example, consider a bivariate $x = (t, u)$ in Examples 1.1 and 1.2. An ANOVA decomposition of a function of x is defined as $\eta(x) = C + g_t + g_u + g_{t,u}$, where C is the constant, g_t and g_u are functions of only one variable called the main effects, and $g_{t,u}$ is the interaction. The decomposition can be made unique by imposing appropriate side conditions on g_t, g_u, and $g_{t,u}$. For regression, setting $g_{t,u} = 0$ results in the so-called additive models; for density estimation, $g_{t,u} = 0$ implies mutual independence of t and u. The structure fits hazard estimation naturally, where forcing $g_{t,u} = 0$ yields proportional hazard models.

The ANOVA decomposition can be built into penalized likelihood estimation using the tensor-product spline technique; see, e.g., Gu and Wahba (1991, 1993). The purpose of this article is to explore the numerical feasibility of automatic estimation of densities and hazards with ANOVA-based structures built in. Algorithms for calculating the estimates with an automatic λ but a completely specified J have been developed in previous work; see Gu (1993a, 1994). With an ANOVA decomposition built in, say $\eta = \sum_\beta g_\beta$ with β being a generic index, however, J is usually of the form $\sum_\beta \theta_\beta^{-1} J_\beta(g_\beta)$, where $J_\beta(g_\beta)$ measures the roughness of g_β, and the weights θ_β, an extra set of smoothing parameters, should naturally also be selected adaptively.

Some related recent work on multivariate density estimation are Stone (1994) and Sain, Baggerly, and Scott (1994). Stone (1994) proposes the use of tensor product regression splines in the context and discusses some theoretical properties, and we look forward to seeing the numerical implementation of the method. Sain et al. (1994) study smoothing parameter selection in multivariate kernel density estimation. Among related work on hazard estimation are Gray (1992) and Kooperberg, Stone and Truong (1995). Gray (1992) experiments with models proposed by Zucker and Karr (1990). Kooperberg et al. (1995) present an implementation of the use of tensor product regression splines in hazard estimation.

The rest of the article is organized as follows. In section 2, background materials are briefly reviewed and the numerical problem is specified. Section 3 describes a feasible automatic multiple smoothing parameter algorithm for the calculation of density and hazard estimates of (1.3) and (1.4). Density estimation and hazard estimation examples are presented in Sections 4 and 5, respectively, to illustrate potential applications of the technique, with notes on numerical and statistical performance of the method. Section 6 concludes the article with some further discussion.

2. Formulation and Preliminaries

In this section, we will discuss a few basic technical facts to tighten up the setup of the problem, present in some detail a specific formulation to be used in later sections, and review some background theoretical and algorithmic results.

For the statistical models implied by (1.1), $L(\eta)$ represents the stochastic part and $\lambda J(\eta)$ the systematic part. The minimization of (1.1) is implicitly over a function space $\mathcal{H} \subseteq \{\eta : J(\eta) < \infty\}$. The penalty $J(\eta)$ forms a natural square (semi) norm in $\mathcal{H}$, and supplemented by a norm in $J_\perp$, makes $\mathcal{H}$ a Hilbert space. Evaluation $\eta(x)$ appears in the $L(\eta)$ part of (1.3) and (1.4). To make the functional $L(\eta) + (\lambda/2)J(\eta)$ continuous in η, it is necessary that evaluation is continuous in $\mathcal{H}$. A Hilbert space in which evaluation is continuous is called a reproducing kernel Hilbert space (RKHS) possessing a reproducing kernel (RK) $R(x, y)$, a positive definite bivariate function satisfying $\langle \eta(\cdot), R(x, \cdot)\rangle = \eta(x)$ (the reproducing property). A mathematical theory of RKHS can be found in Aronszajn (1950); see also Wahba (1990), Chapter 1. The inner-product $\langle \cdot, \cdot \rangle$ (hence norm) and the RK R define each other uniquely. Given a norm in $J_\perp$, $\mathcal{H}_J = \mathcal{H} \ominus J_\perp$ is an RKHS with a square norm J and an RK, say, R_J, and the systematic part of the model implied by (1.1) is effectively determined by $J_\perp$, R_J, and the smoothing parameter λ.

We now specify the construction of an RKHS with an ANOVA decomposition built in on $[0, 1]^2$. Side conditions in the ANOVA decomposition will affect the construction, and in the examples of this article we set $\int g_t dt = \int g_u du = \int g_{t,u} dt = \int g_{t,u} du = 0$. Starting from any positive definite function $R(x, y)$ on a domain $\mathcal{X}$, an inner-product can be defined in $\{R(x, \cdot), x \in \mathcal{X}\}$ to make it an RKHS with $R(x, y)$ as its RK (cf. Aronszajn (1950)), hence it suffices to construct an RK on the domain. The approach of Aronszajn (1950) for RK construction on a product domain starts with the construction of RK's on marginal domains. On the marginal domain $[0, 1]$, a commonly used roughness measure is $J = \int \ddot{\eta}^2 dx$ with $J_\perp = \{1, (\cdot - .5)\}$. The function space $\{g : \int \ddot{g}^2 dx < \infty\}$ can be written as a tensor sum $\mathcal{H}_c \oplus \mathcal{H}_\pi \oplus \mathcal{H}_s$, where $\mathcal{H}_c = \{1\}$ has an RK $R_c = 1$, $\mathcal{H}_\pi = \{(\cdot - .5)\}$ has an RK $R_\pi(t, s) = (t - .5)(s - .5)$, and $\mathcal{H}_s = \{g : \int \ddot{g}^2 dx < \infty, \int g dx = \int \dot{g} dx = 0\}$

has an RK $R_s(t,s) = k_2(t)k_2(s) - k_4(|t-s|)$ dual to the norm $J(g) = \int \ddot{g}^2 dx$, where $k_\nu = B_\nu/\nu!$ and B_ν are the νth Bernoulli polynomials (cf. Craven and Wahba (1979)). A univariate ANOVA decomposition is in place where $\mathcal{H}_c$ carries the constant and $\mathcal{H}_\pi \oplus \mathcal{H}_s$ carries the "treatment effect" satisfying the side condition $\int g dx = 0$. The product of a positive definite function on $\mathcal{T}$ and a positive definite function on $\mathcal{U}$ is a positive definite function on $\mathcal{T} \times \mathcal{U}$ (cf. Aronszajn (1950)), so an RK on a product domain can most conveniently be constructed by taking the product of marginal RK's, and the resulting RKHS is called the tensor product of the corresponding marginal RKHS's. From the three term tensor sum decomposition of the marginal RKHS above, one naturally obtains a tensor product RKHS with nine tensor sum terms $\mathcal{H} = \oplus_{\beta \in \{c,\pi,s\}^2} \mathcal{H}_\beta$, where for example $\mathcal{H}_{s,s}$ is generated from the RK $R_{s,s}((t,u),(s,v)) = R_s(t,s)R_s(u,v)$. An ANOVA decomposition is in place where $\mathcal{H}_{c,c}$ carries the constant, $\mathcal{H}_{\pi,c} \oplus \mathcal{H}_{s,c}$ carries the t main effect, $\mathcal{H}_{c,\pi} \oplus \mathcal{H}_{c,s}$ carries the u main effect, and $\mathcal{H}_{\pi,\pi} \oplus \mathcal{H}_{\pi,s} \oplus \mathcal{H}_{s,\pi} \oplus \mathcal{H}_{s,s}$ carries the interaction. Let the roughness penalty be $J = \sum_\beta \theta_\beta^{-1} J_\beta$ where J_β are the square norm in $\mathcal{H}_\beta$. Setting $\theta_\beta = 0$ eliminates $\mathcal{H}_\beta$ from the model space and setting $\theta_\beta = \infty$ puts $\mathcal{H}_\beta$ in $J_\perp$. One has $\mathcal{H}_J = \oplus_{\theta_\beta \in (0,\infty)} \mathcal{H}_\beta$ and $R_J = \sum_{\theta_\beta < \infty} \theta_\beta R_\beta$. The subspaces $\mathcal{H}_{c,c}$, $\mathcal{H}_{c,\pi}$, $\mathcal{H}_{\pi,c}$, and $\mathcal{H}_{\pi,\pi}$ are of finite dimension and are often included in $J_\perp$. The other terms can only appear in $\mathcal{H}_J$. For density estimation, the constant $\mathcal{H}_{c,c}$ should be eliminated to maintain a one-to-one logistic density transform $f \leftrightarrow e^\eta / \int e^\eta dx$. Formulas of J_β in this construction can be found in Gu (1996) but are not needed for computation.

When $L(\eta)$ depends on η only through evaluations $\eta(X_i)$ as in the regression problem of Example 1.1, the solution of (1.1) is in a data-adaptive finite dimensional subspace $\mathcal{H}_n = J_\perp \oplus \{R_J(X_i, \cdot)\}$ (cf. Wahba (1990)). The restriction to a finite dimensional space makes the numerical calculation of the estimates possible. For density estimation, the minimizer $\hat{\eta}$ of (1.3) in $\mathcal{H}$ generally does not have a finite dimensional expression. Nevertheless, an asymptotic analysis in Gu and Qiu (1993) shows that there is no loss of asymptotic efficiency when the model space is restricted to $\mathcal{H}_n$, in the sense that the minimizer $\hat{\eta}_n$ of (1.3) in $\mathcal{H}_n$ shares the same asymptotic convergence rates as $\hat{\eta}$, so in practice one may calculate $\hat{\eta}_n$ to estimate η.

Write $\xi_i = R_J(X_i, \cdot)$ and $J_\perp = \{\phi_\nu\}_{\nu=1}^M$. A function $\eta \in \mathcal{H}_n$ has an expression $\eta(x) = \sum_{i=1}^n c_i \xi_i(x) + \sum_{\nu=1}^M d_\nu \phi_\nu(x)$. Fixing smoothing parameters, $\hat{\eta}_n$ can be calculated by minimizing

$$-\frac{1}{n}\mathbf{1}^T(Q\boldsymbol{c} + S\boldsymbol{d}) + \log \int \exp(\boldsymbol{\xi}^T\boldsymbol{c} + \boldsymbol{\phi}^T\boldsymbol{d})dx + \frac{\lambda}{2}\boldsymbol{c}^T Q \boldsymbol{c}, \tag{2.1}$$

where $\boldsymbol{\xi}$ and $\boldsymbol{\phi}$ are vectors of functions and $\boldsymbol{c}$ and $\boldsymbol{d}$ are vectors of coefficients, Q is $n \times n$ with (i,j)th entry $R_J(X_i, X_j) = J(\xi_i, \xi_j)$ where $J(\cdot,\cdot)$ indicates the

inner-product in $\mathcal{H}_J$, and S is $n \times M$ with (i,ν)th entry $\phi_\nu(X_i)$. Let $\mu_\eta(h) = \int he^\eta dx / \int e^\eta dx$ and $V_\eta(h,g) = \mu_\eta(hg) - \mu_\eta(h)\mu_\eta(g)$. From an estimate $\tilde\eta = \boldsymbol{\xi}^T\tilde{\boldsymbol{c}} + \boldsymbol{\phi}^T\tilde{\boldsymbol{d}}$, the one-step Newton update for minimizing (2.1) satisfies

$$\begin{pmatrix} V_{\xi,\xi} + \lambda Q & V_{\xi,\phi} \\ V_{\phi,\xi} & V_{\phi,\phi} \end{pmatrix} \begin{pmatrix} \boldsymbol{c} \\ \boldsymbol{d} \end{pmatrix} = \begin{pmatrix} Q\mathbf{1}/n - \boldsymbol{\mu}_\xi + V_{\xi,\eta} \\ S^T\mathbf{1}/n - \boldsymbol{\mu}_\phi + V_{\phi,\eta} \end{pmatrix}, \tag{2.2}$$

where $\boldsymbol{\mu}_\xi = \mu_{\tilde\eta}(\boldsymbol{\xi})$, $\boldsymbol{\mu}_\phi = \mu_{\tilde\eta}(\boldsymbol{\phi})$, $V_{\xi,\xi} = V_{\tilde\eta}(\boldsymbol{\xi},\boldsymbol{\xi}^T)$, $V_{\xi,\phi} = V_{\tilde\eta}(\boldsymbol{\xi},\boldsymbol{\phi}^T)$, $V_{\phi,\phi} = V_{\tilde\eta}(\boldsymbol{\phi},\boldsymbol{\phi}^T)$, $V_{\xi,\eta} = V_{\tilde\eta}(\boldsymbol{\xi},\tilde\eta)$, and $V_{\phi,\eta} = V_{\tilde\eta}(\boldsymbol{\phi},\tilde\eta)$; see Gu (1993a) for details.

With varying smoothing parameters, (2.2) defines a class of estimates, and one may try to choose a better performing one from the class as the update. A performance-oriented iteration simultaneously updating λ and η was developed in Gu (1993a), where the performances of η's are compared on a computable proxy $\hat L_{\tilde\eta}(\eta,\eta_0)$ (cf. Gu (1993a), Eq. 3.6) of the symmetrized Kullback-Leibler between the true density $e^{\eta_0}/\int e^{\eta_0}dx$ and the estimate $e^\eta/\int e^\eta dx$, $\mathrm{SKL}(\eta,\eta_0) = \mu_{\eta_0}(\eta_0-\eta) + \mu_\eta(\eta-\eta_0)$, where the one-step Newton update η is dependent on $\tilde\eta$, θ_β, and λ through the terms of (2.2). The arguments and formulas in Gu (1993a), Section 3 remain valid when θ_β hidden in R_J are also to be updated, but the single smoothing parameter algorithm of Gu (1993a), Section 4 is no longer sufficient. More details and a multiple smoothing parameter algorithm will be described in the next section.

Parallel results hold for hazard estimation. Let $N = \sum_{i^*=1}^n \delta_{i^*}$ and T_i, $i = 1,\ldots,N$, be the observed failure times, where i^* runs over all observations but i only runs over observed failures. The minimizer $\hat\eta_n$ of (1.4) in $\mathcal{H}_n = J_\perp \cup \{R_J((T_i,U_i),\cdot)\}$ shares the same asymptotic convergence rates as $\hat\eta$ in $\mathcal{H}$ (cf. Gu (1996)). The estimate $\hat\eta_n = \sum_{i=1}^N c_i R_J((T_i,U_i),\cdot) + \sum_{\nu=1}^M d_\nu\phi_\nu(\cdot) = \boldsymbol{\xi}^T\boldsymbol{c} + \boldsymbol{\phi}^T\boldsymbol{d}$ can be computed via minimizing

$$-\frac{1}{n}\mathbf{1}^T(Q\boldsymbol{c} + S\boldsymbol{d}) + \frac{1}{n}\sum_{i^*=1}^n \int_{\mathcal{T}} Y_{i^*}\exp(\boldsymbol{\xi}_{i^*}^T\boldsymbol{c} + \boldsymbol{\phi}_{i^*}^T\boldsymbol{d})dt + \frac{\lambda}{2}\boldsymbol{c}^TQ\boldsymbol{c}, \tag{2.3}$$

where Q is $N\times N$ with (i,j)th entry $\xi_i(T_j,U_j) = R_J((T_i,U_i),(T_j,U_j))$, S is $N\times M$ with (i,ν)th entry $\phi_\nu(T_i,U_i)$, $Y_{i^*}(t) = I_{[X_{i^*}\ge t > Z_{i^*}]}$ is the at-risk process of the i^*th observation, $\boldsymbol{\xi}_{i^*}$ is $N\times 1$ with ith entry $\xi_i(t,U_{i^*})$, and $\boldsymbol{\phi}_{i^*}$ is $M\times 1$ with νth entry $\phi_\nu(t,U_{i^*})$. The one-step Newton update for minimizing (1.4) again satisfies (2.2) but with the entries modified according to the modified definitions of $\mu_\eta(h) = (1/n)\sum_{i^*=1}^n \int_{\mathcal{T}} h_{i^*}Y_{i^*}e^{\eta_{i^*}}dt$ and $V_\eta(h,g) = \mu_\eta(hg)$, where $h_{i^*}(t) = h(t,U_{i^*})$ and $\eta_{i^*}(t) = \eta(t,U_{i^*})$. With a performance measure $\mathrm{SKL}(\eta,\eta_0) = \int_{\mathcal{U}}\int_{\mathcal{T}}(e^\eta - e^{\eta_0})(\eta-\eta_0)\tilde S m dt du$ where $\tilde S(t,u) = P(X\ge t > Z|U=u)$ is the at-risk probability and $m(u)$ is the density of U, the formula of $\hat L_{\tilde\eta}(\eta,\eta_0)$ holds verbatim, up to the entries appearing in (2.2), as a computable performance proxy for the one-step

Newton updates in the hazard estimation setting. An argument can be found in Gu (1994) for a singleton $\mathcal{U}$, which extends readily to the general setting.

Finally come a few words on the uniqueness of the estimate. Fixing smoothing parameters, (1.3) is strictly convex and has a unique solution in $\mathcal{H}_n$ as long as the maximum likelihood estimate exists in $J_\perp$ (cf. Gu and Qiu (1993)). For (1.4), an extra condition is needed for strict convexity, that $\mathcal{H}_n$ keeps its dimension on the restricted domain $\cup_{i=1}^n\{(Z_i, X_i] \times \{U_i\}\}$ (cf. Gu (1996)). This is usually not a problem in practice. Even when (1.3) or (1.4) has a unique solution, (2.1) or (2.3) may not have a unique minimizing $\mathbf{c}$ if there are replicated data which yield duplicated ξ_i, but the algorithm is carefully designed to take care of this; see discussion in Gu (1993a), Appendix.

3. Algorithm

Write $V_\eta(h) = V_\eta(h, h)$. One has $\mathrm{SKL}(\eta, \eta_0) = V_{\eta'}(\eta - \eta_0)$ and $\mu_{\tilde\eta}(\eta) - \mu_{\eta_0}(\eta) = V_{\eta''}(\eta, \tilde\eta - \eta_0)$, where η' is a convex combination of η and η_0, and η'' that of $\tilde\eta$ and η_0. Replacing η' and η'' by $\tilde\eta$ in the preceding equations, one can derive a proxy of $\mathrm{SKL}(\eta, \eta_0)$ of the form $A(\eta, \tilde\eta) - 2\mu_{\eta_0}(\eta) + C(\tilde\eta, \eta_0)$, where $A(\eta, \tilde\eta)$ can be computed, $C(\tilde\eta, \eta_0)$ can be dropped for comparative purposes, and the terms of $\mu_{\eta_0}(\eta)$ can be estimated by sample means or cross-validated sample means. The computable proxy $\hat L_{\tilde\eta}(\eta, \eta_0)$ for the comparison of the one-step Newton updates from $\tilde\eta$ is simply $A(\eta, \tilde\eta)/2 - \hat\mu_{\eta_0}(\eta)$. See Gu (1993a), Section 3 for further details and the formulas for computing $\hat L_{\tilde\eta}(\eta, \eta_0)$.

In a performance-oriented iteration, one tries to minimize $\hat L_{\tilde\eta}(\eta, \eta_0)$ with respect to the smoothing parameters where η is the one-step Newton update from $\tilde\eta$ satisfying (2.2), and one takes the resulting η as the new $\tilde\eta$. When the iteration converges to a fixed point, the converged η is clearly the minimizer $\hat\eta_n$ of (2.1) or (2.3) with the smoothing parameters set to the converged values, and there is no other one-step Newton update that performs better according to the estimated performance measure $\hat L_{\tilde\eta}(\eta, \eta_0)$. Informally, the existence of the fixed point and the convergence of the iteration depends on how parallel $\hat L_{\tilde\eta}(\eta, \eta_0)$ of different $\tilde\eta$'s are to each other, and the performance of the converged estimate depends on how parallel the $\hat L_{\tilde\eta}(\eta, \eta_0)$ at convergence is to $\mathrm{SKL}(\eta, \eta_0)$, as functions of the smoothing parameters; an analytical treatment seems formidable.

The multiple smoothing parameter algorithm we will be using consists of an initialization step and an updating step. For the initial value of θ_β, say θ_1, one first sets $\theta_1 = 1$ and $\theta_\gamma = 0$, $\gamma \neq 1$, then invokes the single smoothing parameter algorithm of Gu (1993a), Section 4 to obtain an automatic λ, say λ_1, with R_1 being the only penalized term, and then sets $\theta_1 = 1/\lambda_1$. After separate calculations of initial θ_β in this manner, one puts all penalized terms back together and employs the fixed θ_β algorithm once more to set up for the

updating step. Such a procedure is invariant of individual scalings of R_β, which are usually arbitrary and not comparable to each other. When the penalized terms contribute somewhat "independently" to the estimate, in the sense that each term's relevance/importance are not affected much by the presence of other terms in the estimate, the relative weights chosen this way should not be too far from the "optimal" ones.

In each iteration of the updating step, one first fixes λ and $\tilde{\eta}$ and updates θ_β one at a time through the list, then invokes the fixed θ_β algorithm to calculate λ and $\tilde{\eta}$ for the next iteration. To update a certain θ_β, say θ_1, other θ_β are fixed at their latest values and $\hat{L}_{\tilde{\eta}}(\eta, \eta_0)$ is evaluated at three different values of θ_1: the current value, say $\tilde{\theta}_1$, and two adjacent values $\tilde{\theta}_1 10^{\pm .1}$; a quadratic in $\log_{10} \theta_1$ is then fit through the three points and the minimum of the quadratic on $[\log_{10} \tilde{\theta}_1 - .5, \log_{10} \tilde{\theta}_1 + .5]$ is determined, and $\hat{L}_{\tilde{\eta}}(\eta, \eta_0)$ is evaluated once more at the minimum; the smallest of the four evaluations gives the new θ_1. Note that this is just a standard safe-guarded Newton search, with the first and second derivatives at $\tilde{\theta}_1$ approximated by finite differences. The order in which θ_β is updated could be arbitrary, but for definiteness we choose to follow the descending order of the traces of $(\theta_\beta R_\beta(X_i, X_j))$ at the outset of each iteration. Note that only relative values of θ_β matter so a standardization procedure should follow the θ_β updating, for which we choose to set the trace of Q (cf. 2.2) to one. The algorithm is clearly invariant of the scaling and the indexing of R_β. Convergence is declared when the supremum change in $e^{\tilde{\eta}}$ is within a user-supplied error limit, where the supremum is taken over evaluations on the quadrature points and the data points that contribute to the terms of (2.2), and the change is measured by a combination of the absolute and relative error similar to the suggestion of Gill, Murray and Wright (1981), Section 2.1.1. For the "inner-loop" fixed θ_β iteration, an error limit in the order of $10^{-5} \sim 10^{-6}$ is stringent yet affordable; for the "outer loop" of the iteration, an error limit in the order of $10^{-2.5} \sim 10^{-3}$ works well in practice.

The choice of such a simple coordinate-wise updating procedure is out of the following considerations. First, the derivatives of $\hat{L}_{\tilde{\eta}}(\eta, \eta_0)$ with respect to the smoothing parameters are beyond reach so the Newton method is not feasible for the minimization of $\hat{L}_{\tilde{\eta}}(\eta, \eta_0)$. Second, $\hat{L}_{\tilde{\eta}}(\eta, \eta_0)$ will change from iteration to iteration anyway and so will the minimizing smoothing parameters, and it is not advisable to invest too much for the minimization of $\hat{L}_{\tilde{\eta}}(\eta, \eta_0)$ in any single iteration; this rules out the usual quasi-Newton approach because the evaluation of $\hat{L}_{\tilde{\eta}}(\eta, \eta_0)$ is costly. Drawing a very crude parallel to a standard optimization method, the one-step Newton updates of (2.2) define a "search direction" whereas the performance proxy $\hat{L}_{\tilde{\eta}}(\eta, \eta_0)$ drives a "line search" on the "direction". Note that our real objective function $\mathrm{SKL}(\eta, \eta_0)$ is beyond reach. We give up on an

exact "line search" for its formidable cost, but the algorithm can still be effective as long as the fixed point of the iteration remains the same.

Numerical details are not of primary interest here and hence are omitted. The algorithm is implemented in portable RATFOR code packaged in RKPACK-II, currently available in beta version at `http://www.stat.purdue.edu/~chong/software.html`. The numerical performance of the algorithm will be discussed along with the examples in sections to follow. Empirical results concerning the statistical performance of the fixed point of the performance-oriented iteration relative to the best possible performance of all $\hat{\eta}_n$ can be found in Gu (1993a, 1994) in settings with a single smoothing parameter. Similar statistical performance is expected of the method in the multiple smoothing parameter settings since the statistical aspects of the method essentially remain the same, but similar simulation studies appear formidable due to the cost of locating the best-performing $\hat{\eta}_n$ with a multivariate index. Simulations of limited scale are included in the next two sections to illustrate the absolute performance of the method.

4. Density Estimation Examples

First let us look at a data set listed in Wang (1989) concerning AIDS patients infected by blood-transfusion. The variables are the time T from HIV infection to AIDS diagnosis and the time U from HIV infection to the end of data collection, both in months. The data set consists of 3 subsets: 34 "children" of age 1–4, 120 "adults" of age 5–59, and 141 "elderly patients" of age 60 or above. Clearly only data with $T \le U$ can be observed, i.e., the observations are randomly truncated. Of interest is the estimation of the distributions of T and U.

Under the assumption of pre-truncation independence of T and U, the distributions of T and U were estimated separately in Gu (1993b) via a penalized conditional likelihood approach. Using penalized full likelihood with multiple smoothing parameters, one may also estimate the distributions of T and U simultaneously. As an illustration, we conduct the analysis for the "elderly patients" and compare the results with those in Gu (1993b).

The pre-truncation domain was chosen to be $[0,100]^2$ which covered all the observations. The domain to use in (1.3) was the triangle $\mathcal{X} = [0,100]^2 \cap \{t \le u\}$. The domain $[0,100]^2$ was mapped onto $[0,1]^2$. Employing the tensor product spline construction of Section 2, we used a null space $J_\perp = \{\phi_1, \phi_2, \phi_3\} = \{(t-.5), (u-.5), (t-.5)(u-.5)\}$ with $M = 3$ dimensions and an RK $R_J = \theta_{s,c}R_{s,c} + \theta_{c,s}R_{c,s} + \theta_{s,\pi}R_{s,\pi} + \theta_{\pi,s}R_{\pi,s} + \theta_{s,s}R_{s,s}$ with 5 terms. Letting $x = (t,u)$, the fit $\eta(x) = \sum_{\nu=1}^{3}\phi_\nu(x)d_\nu + \sum_{i=1}^{n}R_J(X_i,x)c_i$ decomposes into $\eta(x) = g_t(t) + g_u(u) + g_{t,u}(t,u)$, where $g_t = d_1\phi_1(x) + \sum_{i=1}^{n}\theta_{s,c}R_{s,c}(X_i,x)c_i$, $g_u = d_2\phi_2(x) + \sum_{i=1}^{n}\theta_{c,s}R_{c,s}(X_i,x)c_i$, and $g_{t,u} = d_3\phi_3(x) + \sum_{i=1}^{n}(\theta_{s,\pi}R_{s,\pi}(X_i,x) +$

$\theta_{\pi,s}R_{\pi,s}(X_i,x)+\theta_{s,s}R_{s,s}(X_i,x))c_i$. Pre-truncation independence is characterized by $g_{t,u}=0$.

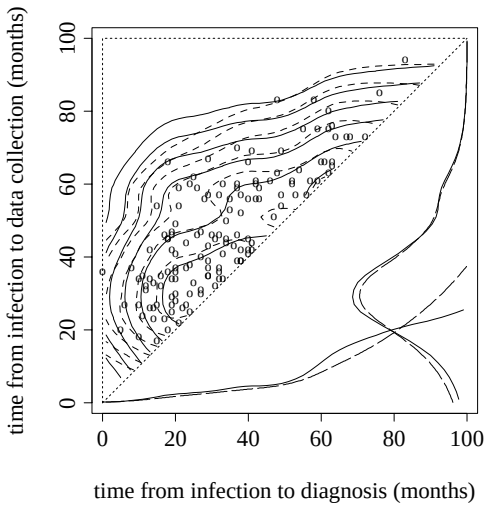

Figure 4.1. Blood-Transfusion Data of "Elderly Patients". Dashed contours are density estimate without pre-truncation independence. Solid contours are density estimate with pre-truncation independence. Marginal densities of the independence model are plotted as solid lines on their axes; marginals estimated by penalized conditional likelihood are superimposed as long-dashed lines.

The estimated density with pre-truncation independence is contoured in Figure 4.1 as solid lines, the data are superimposed as circles, and the domain $\mathcal{X}$ is surrounded by the dotted lines. The integrations were calculated by summation over a 50×50 equally spaced grid restricted to the triangle domain, with the grid points on the diagonal carrying half weight. The marginal densities are superimposed in the blank space on their corresponding axes, where the solid lines are estimates based on full likelihood calculated here and the long-dashed lines are estimates based on conditional likelihoods taken from Gu (1993b). The heights of the solid and long-dashed lines are adjusted so the areas under them are the same. It is clear that the two sets of estimates of the marginals agree well in the light truncation areas but depart a bit in the heavy truncation areas.

The density estimate without pre-truncation independence is also superimposed in Figure 4.1 as dashed lines. To assess the feasibility of pre-truncation independence, one possible diagnostic is the log likelihood ratio $\sum_{i=1}^{141}\log(f_1(X_i)/f_2(X_i))$ where f_1 and f_2 are the estimates without and with pre-truncation independence. In lack of parametric model assumptions for the

systematic part, however, such a score no longer follows the usual χ^2 sampling distribution under the "null", and due to the automatic smoothing parameter selection, such a score may even turn out to be negative occasionally. Nevertheless, the score is informative when it is "too small", say less than $1.9 < \chi^2_{.05,1}/2$, or "too big", say larger than $10 > \chi^2_{.05,10}/2$, where $\chi^2_{\alpha,\nu}$ is the $(1-\alpha)$th percentile of χ^2_ν. The score, however, had value 7.243 in the grey area for the current example, so other means was needed. A closer look at the components of the estimate revealed that $d_3\phi_3$ dominated the interaction term in $\log f_1$, and after calculating a new fit f_3 by setting $d_3 = 0$ and allowing only penalized interaction terms in the estimation, we observed $\sum_{i=1}^{141} \log(f_1(X_i)/f_3(X_i)) = .179$, so the interaction appeared "real" albeit of minor magnitude. One may conclude that the independence model bears a slight lack of fit.

The performance-oriented iteration implemented in the algorithm operates on mathematically different performance proxies $\hat{L}_{\tilde{\eta}}(\eta, \eta_0)$ at different $\tilde{\eta}$, so convergence is not guaranteed. Nevertheless, divergence rarely occurs in our experiments with a single smoothing parameter (cf. Gu (1993a, 1993b, 1994)). Multiple smoothing parameters introduce greater flexibility, and one may expect a bit more difficulties. For the blood-transfusion data, however, the algorithm converged in all cases without incidence. On an IBM-RS6000, a run for the $n = 141$ blood-transfusion example with 5 θ_β's took about 30 cpu minutes.

To empirically check the statistical performance of the estimation tools proposed, we conducted a small scale simulation study, which however is computationally expensive. Truncated data $X = (T, U)$ were generated on $[0,1]^2 \cap \{t \le u\}$ with $T \sim f_0$ and $U \sim f_0$, independent of each other before truncation, where f_0 is the half-half mixture of $N(.3, .01)$ and $N(.7, .01)$ truncated to $[0,1]$. One hundred replicates of samples of size $n = 200$ were generated. Automatic fits were calculated for each of the replicates under two different formulations as in the blood-transfusion example, and the log likelihood $\sum_{i=1}^{200} \log \hat{f}(X_i)$ and the symmetrized Kullback-Leibler $\mathrm{SKL}(\hat{\eta}, \eta_0)$ were collected for each of the fits.

Without pre-truncation independence, call it the dependence model, the iteration converged within 15 outer-loop steps on 90 of the replicates. For the 10 cases on which the iteration did not converge, estimates were still available from the last iterates, and it is not surprising to notice that these fits are rather poor as indicated by the large SKL values; see Figure 4.2 below. With pre-truncation independence, call it the independence model, the iteration converged within 15 outer-loop steps on 87 of the replicates, of which 86 are among the 90 replicates on which the dependence model iteration converged.

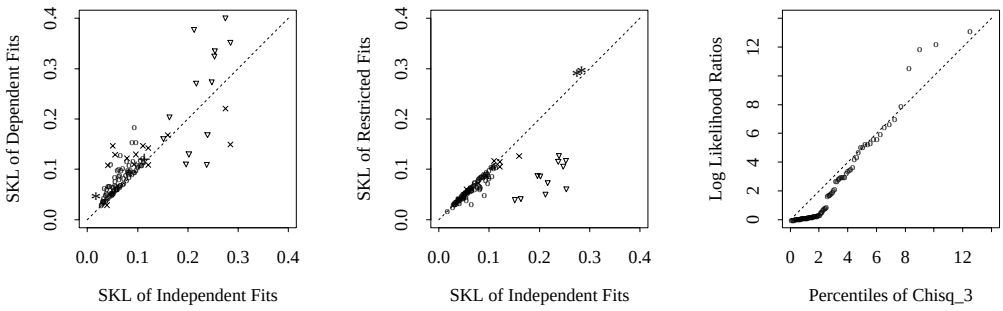

Figure 4.2. Summary of Simulation Results. Left: SKL's of independent and dependent fits. Center: SKL's of unrestricted and restricted independent fits. Right: A "Q-Q plot" of log likelihood ratio.

The diverged cases are virtually trapped to interpolation at certain θ_β combinations. When this happens, however, one still has the choice of using other θ_β's, say the starting values, unless the inner iteration diverges on the outset at the starting θ_β values. Of the 10 diverged dependence replicates, none was stuck at the starting θ_β values. Of the 13 diverged independence replicates, only 2 were stuck at the starting θ_β values. We also calculated fits with starting θ_β values for the diverged cases, call them backup fits, and recorded the corresponding SKL's.

The SKL's of the fits under the two models are plotted against each other in the left frame of Figure 4.2, where circles are cases on which both iterations converged and triangles are cases on which at least one iteration did not converge. The backup fits for diverged cases are also superimposed as crosses. The star marks the best fit under the independence model and the plus marks the worst of the converged fits under the independence model. The best and the worst of the converged fits under the independence model are contoured in the two frames of Figure 4.3 as solid lines, with the test density superimposed as dotted lines and data as circles; the estimated and test marginals are plotted in the blank space as solid and dotted lines, respectively.

Besides the unrestricted independent fits with free θ_1 and θ_2, we also calculated restricted independent fits with $\theta_1 = \theta_2$. The SKL's of the restricted independent fits are plotted against those of the unrestricted fits in the center frame of Figure 4.2, where the circles mark cases on which both iterations converged, triangles mark cases on which the unrestricted iteration did not converge, and the stars mark cases on which both of the iterations did not converge; the backup fits for the diverged cases are superimposed as crosses. The two test marginals are identical, so it is reassuring to see that the restriction leads to better estimates. Such practice is only practical on simulated data, but the exercise is indicative of the fact that the incorporation of extra information often yields performance improvement.

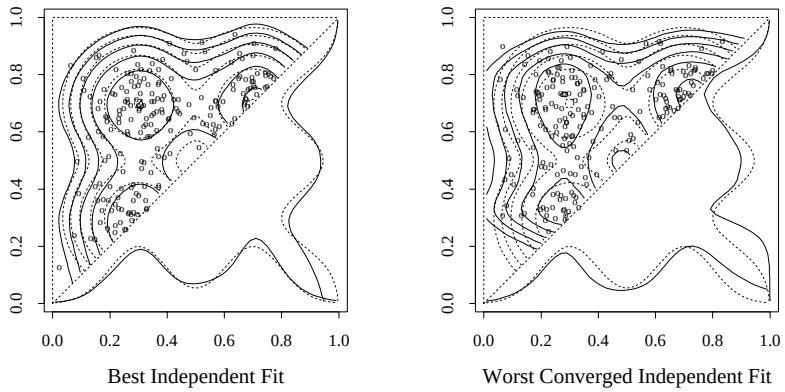

Figure 4.3. Best and Worst Independent Fits. Density estimates are in solid contours and test density in dotted contours. Circles are data. Estimated and test marginal densities are plotted as solid and dotted lines on their axes, respectively.

The right frame of Figure 4.2 plots the ordered log likelihood ratio of the 86 "good" cases, say $l_{(i)}$, versus the $(i-.5)/86\times 100\%$-th percentile of χ^2_3. Note that we are *not* plotting the "χ^2-statistic" $2l_{(i)}$ but one-half of it. Several similar plots with different multiples of the log likelihood ratio and different degrees of freedom for the χ^2 percentiles have been looked at, and the one presented appears to be the "closest" to the 45 degree line. The first 5 $l_{(i)}$'s are negative but very close to 0. It is clear that the log likelihood ratio does *not* follow a χ^2 distribution. We do not yet know how to properly calibrate it, especially in the gray area. Hypothesis testing with a "nonparametric null" has not been well formulated yet, but deserves major research effort.

5. Hazard Estimation Examples

The first example we will be looking at is the Stanford heart transplant data listed in Miller and Halpern (1982). Recorded were survival or censoring times of 184 patients after (the first) heart transplant (in days), their ages at transplant, and a certain tissue type mismatch scores for 157 of the patients. There were 113 recorded deaths and 71 censorings. There is no truncation in the data, i.e., $Z_i \equiv 0$. Due to the insignificance in the analyses by Miller and Halpern (1982) and by others, and also due to the missing values, we first discard the tissue type mismatch score in the analysis.

Let T be time after transplant and U be age at transplant. The time axis was transformed by $t^* = t^{1/2}$ to make the survival/censoring times more evenly scattered, and then the hazard was estimated on $(t^*, u) \in [0, 61] \times [10, 65] = \mathcal{T}^* \times \mathcal{U}$

which covered all the observations. From the estimated hazard on the transformed time axis $e^{\eta(t^*,u)} = -d\log S(t^*,u)/dt^*$, the hazard on the original time axis is simply $e^{\eta(t^*,u)}(dt^*/dt) = e^{\eta(t^{1/2},u)}/(2t^{1/2})$. The domain $\mathcal{T}^*\times\mathcal{U}$ was mapped onto $[0,1]^2$ for calculation using the same tensor product spline construction as in the density estimation examples but with the constant term included.

The fitted $e^{\eta(t^*,u)}$ with interaction is contoured as dashed lines and that without interaction (proportional hazard, PH henceforth) as solid lines in the left frame of Figure 5.1, where the data are superimposed as circles (deceased) or crosses (censored). To assess the plausibility of hazard proportionality, the log likelihood ratio $\sum_{i=1}^{184}\{\delta_i(\eta_2-\eta_1)(X_i,U_i)-\int_0^{X_i}(e^{\eta_2(t,U_i)}-e^{\eta_1(t,U_i)})dt\}$ (cf. 1.4) was calculated to be 3.376, which led to a "χ^2-statistic" 6.65, where η_1 is the PH fit and η_2 is the fit with interaction. Once again we are in the gray area. Looking at the "p-values" .010, .036, and .084 of 6.65 treated as χ^2_ν, $\nu = 1,2,3$, one however may conclude that the interaction is at most marginally significant, and in turn hazard proportionality looks plausible. The "base hazard" on the original time axis $e^{g_\emptyset+g_{t^*}(t^{1/2})}/(2t^{1/2})$ and the age multiplier e^{g_u} in the proportional hazard model share the right frame of Figure 5.1 as the solid line on the solid axes and the dotted line on the dotted axes, respectively. It can be seen that beyond the first 250 days or so highly hazardous period the risk remains rather stable through the rest of the time axis, with the lowest risk at about 750 days after transplant. The age effect is virtually uniform for those under 40 but the risk takes off quickly beyond age 45.

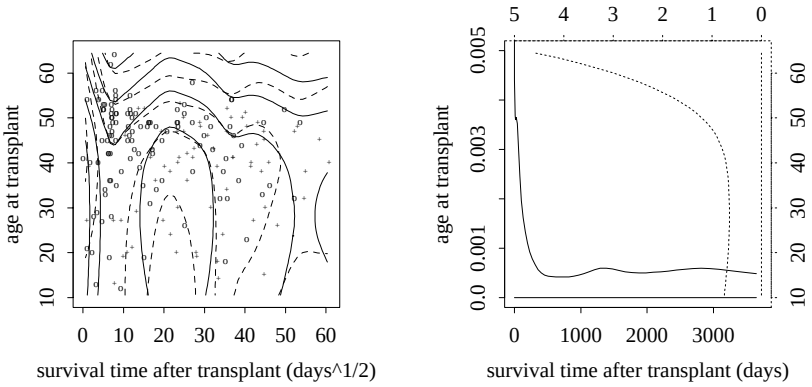

Figure 5.1. Stanford Heart Transplant Data. Left: $e^{\eta(t^*,u)}$ without interaction (solid lines) and with interaction (dashed lines); circles are observed deaths and crosses censorings. Right: "Base hazard" on the original time axis $e^{g_\emptyset+g_{t^*}(t^{1/2})}/(2t^{1/2})$ and age multiplier e^{g_u} of the proportional hazard fit.

To double check the relevance of tissue type mismatch scores, we also tried a few proportional hazard models with covariates $U = (V,W) = (\text{age}, \text{mismatch})$

using the 157 complete data points. Fitted were models of forms $\eta_1(t^*, v, w) = g_\emptyset + g_{t^*} + g_v$, $\eta_2(t^*, v, w) = g_\emptyset + g_{t^*} + g_v + g_w$, and $\eta_3(t^*, v, w) = g_\emptyset + g_{t^*} + g_v + g_w + g_{v,w}$. The log likelihood ratios were $\sum_{i=1}^{157}\{\delta_i(\eta_2-\eta_1)(X_i, U_i) - \int_0^{X_i}(e^{\eta_2(t,U_i)} - e^{\eta_1(t,U_i)})dt\} = .348$ and $\sum_{i=1}^{157}\{\delta_i(\eta_3 - \eta_2)(X_i, U_i) - \int_0^{X_i}(e^{\eta_3(t,U_i)} - e^{\eta_2(t,U_i)})dt\} = 1.012$, so it appeared appropriate to discard the tissue type mismatch scores. Detailed formulations are not of primary interest and hence are omitted. See, e.g., Gu and Wahba (1993) for a general discussion concerning the construction of tensor-product splines.

All iterations in the Stanford heart transplant data example converged without incidence. Entries of (2.2) for hazard estimation have multiple terms of integrals as seen in (2.3), so the calculation is generally slower than that for density estimation. For example, the $n = 184$ fit with interaction took 50 cpu minutes on an IBM-RS6000 and the $n = 184$ proportional hazard fit took 29.

We now conduct a simulation study to assess the performance of the technique for hazard estimation. Data of size $n = 200$ were generated, where we took $U_{4(j-1)+k} = (j - .5)/50$, $j = 1, \ldots, 50$, $k = 1, \ldots, 4$. The life time $T|U$ was generated from the distribution with a proportional hazard function $e^{\eta_0} = 24t^2 \exp(4(u - .5)^2)$, the censoring time C from an exponential density $e^{-c/3}$, and the truncation time Z from an exponential density e^{-5z}, independent of each other. Using the same tensor-product spline construction, automatic fits with and without interaction were calculated on 100 replicates. The log likelihood $\sum_{i=1}^{200}\{\delta_i\hat{\eta}(X_i, U_i) - \int_{Z_i}^{X_i}(e^{\hat{\eta}(t,U_i)})\}$, the symmetrized Kullback-Leibler $\mathrm{SKL}(\hat{\eta} - \eta_0) = \int_{\mathcal{U}}\int_{\mathcal{T}}(e^{\hat{\eta}} - e^{\eta_0})(\hat{\eta} - \eta_0)\tilde{S}m$, and a weighted mean square error $\mathrm{MSE}(\hat{\eta} - \eta_0) = \int_{\mathcal{U}}\int_{\mathcal{T}}(\hat{\eta} - \eta_0)^2 e^{\eta_0}\tilde{S}m$ were collected for all fits, where $\tilde{S}(t, u) = P(X \geq t > Z|U = u)$ and $m(u)$ were substituted by their respective empirical versions.

For the PH model, the iteration converged within 15 outer-loop steps on 98 replicates and diverged towards interpolation on the other 2. With interaction, the iteration converged within 15 outer-loop steps on 98 replicates, diverged towards interpolation on 1 replicate, and was zigzagging after 15 outer-loop steps on 1 replicate.

Plotted in the left and center frames of Figure 5.2 are the SKL's and MSE's of the PH fits versus those of the fits with interaction, where the circles, the star, and the plus are 97 "good" cases and the triangle is the case with converged PH fit but zigzagging interaction fit. The 2 diverged PH fits are not included. The best and the worst PH fits marked by the star and the plus in the left and center frames of Figure 5.2 are plotted in the two frames of Figure 5.3 in the same manner as in the right frame of Figure 5.1. Similar to the right frame of Figure 4.3, the right frame of Figure 5.2 shows plots of the log likelihood ratios versus the $(i - .5)/98 \times 100$th-percentiles of χ^2_6 for the 98 cases plotted in the left and

center frames, where 7 negative log likelihood ratios are truncated to 0. Again the log likelihood ratio does not follow a χ^2 distribution, and inference can be drawn only for "extreme" values.

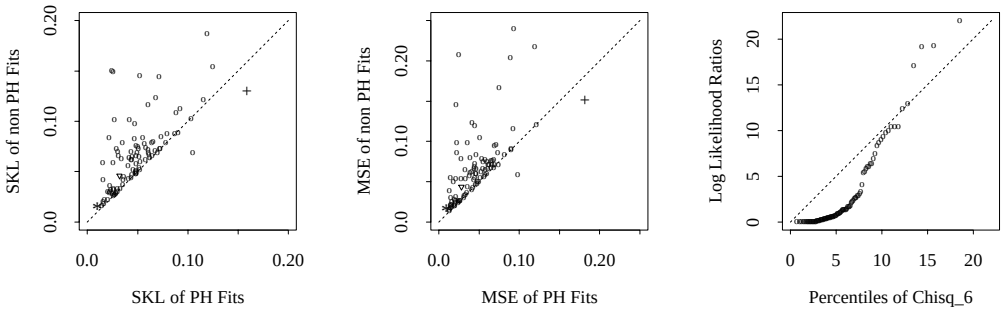

Figure 5.2. Summary of Simulation Results. Left: SKL's of PH fits and fits with interaction. Center: MSE's of PH fits and fits with interaction. Right: A "Q-Q plot" of log likelihood ratio.

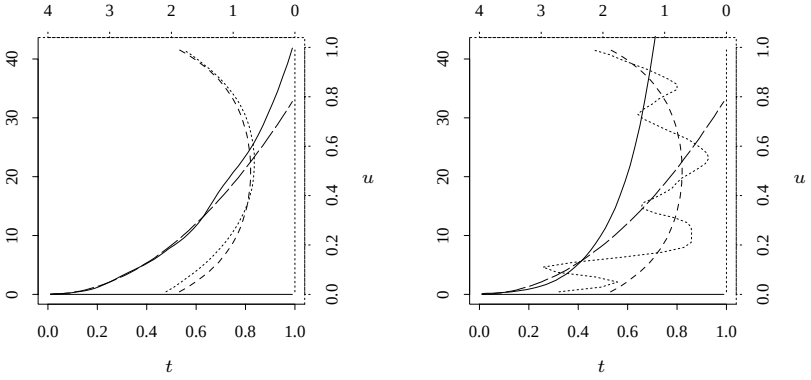

Figure 5.3. Best and Worst PH Fits. Estimates of base hazard are in solid lines on solid axes and those of covariate effect in dotted lines on dotted axes. Corresponding components of test hazard are in long and short dashed lines.

6. Discussion

In this article, structural nonparametric estimation of multivariate probability densities and covariate dependent hazard functions is implemented through tensor product splines. Examples are presented to illustrate potential applications of the technique in data analysis. Although demanding in memory and execution time, the algorithm is generic to fit various model configurations and the data-driven multiple smoothing parameter selection makes the estimation fully

automatic. The code comprises part of a collection of RATFOR routines for penalized likelihood density and hazard estimation by the name RKPACK-II, available in beta version at `http://www.stat.purdue.edu/~chong/software.html`, as a sequel to RKPACK, a collection of routines for smoothing spline regression.

In the existing literature on multivariate nonparametric density estimation, little attention is paid to the exploration/exploitation of independence structures of random variables and no means seems available to allow for truncated domains. The blood-transfusion example of Section 4 shows how these aspects may be incorporated in estimation using tensor product splines. Tensor product estimate respects qualitatively different axes and the automatic selection of smoothing parameters makes the estimation invariant with respect to axis scaling. On multidimensional domains with comparable scaling but not so interpretable axes such as geographical maps, rotation invariant estimation using thin-plate splines would be more appropriate.

Generalizations of Cox's (1972) partial likelihood proportional hazard model have received much attention in recent literature. Cast as special cases of models available through tensor product splines, O'Sullivan (1988b) set $g_{t,u} = 0$ while Zucker and Karr (1990) restricted $g_u \in \mathcal{H}_{c,\pi}$ and $g_{t,u} \in \mathcal{H}_{\pi,\pi} \oplus \mathcal{H}_{s,\pi}$, and both treated the "base hazard" $e^{g_\emptyset + g_t}$ as nuisance and employed penalized partial likelihood to estimate the remaining terms. Gray (1992) illustrated the use of regression splines in penalized partial likelihood for fitting these models, with the amount of smoothing tuned via a certain definition of "degrees of freedom". In comparison, all terms are estimated simultaneously via penalized full likelihood in this article, and the amount of smoothing is tuned automatically according to the estimated performance proxies of the fits. Kooperberg et al. (1995) implemented an adaptive tensor product linear regression spline approach to the estimation of covariate dependent hazard functions, where the ANOVA decomposition is implicit.

For density estimation, the development represents a modest step forward towards nonparametric estimation of graphical models (cf. Whittaker (1990)). Related work on conditional density estimation can be found in Gu (1995). For hazard estimation, a further topic is the incorporation of a time dependent covariate, on which a theory has yet to be developed.

If the observed (T, U) center around some monotone curve in $\mathcal{T} \times \mathcal{U}$, the estimated terms in an ANOVA decomposition may suffer from an identifiability problem called concurvity. If T and U are independent for distribution data or if hazard proportionality holds for survival data, the estimated interaction in an ANOVA decomposition should be negligible. This calls for the assessment of the practical "significance" of the estimated ANOVA terms; some informal analysis of this nature has been explored in the examples, such as the use of

the log likelihood ratio. A systematic treatment of these issues however poses a challenging problem by itself, and is beyond the scope of this article. Some diagnostic tools for regression can be found in Gu (1992), which may help to provide heuristics for similar development in density and hazard estimation.

Numerically, the execution time for each iteration of the algorithm is of the order $O(kn^3)$, where k is the number of θ_β's and n is the number of observations. This poses practical limitations on the size and complexity of the problems that can be feasibly solved using the techniques developed. From a statistical point of view, the larger the k is the less reliable $\hat{L}_{\tilde{\eta}}(\eta, \eta_0)$ tend to be as proxies of $\mathrm{SKL}(\eta, \eta_0)$ (cf. Section 3), so one has to impose some structures for high dimensional problems; this is somewhat related to the very curse of dimensionality. For very large n, one may calculate penalized likelihood estimates of the form $\eta = \sum_{i=1}^{m} c_i \xi_i + \sum_{\nu=1}^{M} d_\nu \phi_\nu$ with execution time of order $O(km^3)$, where $\{\xi_i,\ i = 1, \ldots, m\}$ is a subset of $\{R_J(X_i, \cdot),\ i = 1, \ldots, n\}$. Details await further study.

Acknowledgement

This research was supported by grants from National Science Foundation and from Purdue Research Foundation.

References

Aronszajn, N. (1950). Theory of reproducing kernels. *Trans. Amer. Math. Soc.* **68**, 337-404.

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

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.* **31**, 377-403.

Gill, P. E., Murray, W. and Wright, M. H. (1981). *Practical Optimization*, Academic Press, London.

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

Gray, R. J. (1992). Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *J. Amer. Statist. Assoc.* **87**, 942-951.

Gu, C. (1992). Diagnostics for nonparametric regression models with additive terms. *J. Amer. Statist. Assoc.* **87**, 1051-1058.

Gu, C. (1993a). Smoothing spline density estimation: A dimensionless automatic algorithm. *J. Amer. Statist. Assoc.* **88**, 495-504.

Gu, C. (1993b). Smoothing spline density estimation: Biased sampling and random truncation. Technical Report 92-03 (Rev.), Purdue University, Dept. of Statistics.

Gu, C. (1994). Penalized likelihood hazard estimation: Algorithm and examples. In *Statistical Decision Theory and Related Topics V* (Edited by S. S. Gupta and J. Berger), 61-72. Springer-Verlag.

Gu, C. (1995). Smoothing spline density estimation: Conditional distribution. *Statist. Sinica* **5**, 709-726.

Gu, C. (1996). Penalized likelihood hazard estimation: A general procedure. *Statist. Sinica* **6**, 861-876.

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

Gu, C. and Wahba, G. (1991). Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method. *SIAM J. Sci. Statist. Comput.* **12**, 383-398.

Gu, C. and Wahba, G. (1993). Smoothing spline ANOVA with component-wise Bayesian confidence intervals. *J. Comput. Graphical Statist.* **2**, 97-117.

Kooperberg, C., Stone, C. J. and Truong, Y. K. (1995). Hazard regression. *J. Amer. Statist. Assoc.* **90**, 78-94.

Leonard, T. (1978). Density estimation, stochastic processes and prior information (with discussion). *J. Roy. Statist. Soc. Ser. B* **40**, 113-146.

Miller, R. and Halpern, J. (1982). Regression with censored data. *Biometrika* **69**, 521-531.

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

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

O'Sullivan, F., Yandell, B. and Raynor, W. (1986). Automatic smoothing of regression functions in generalized linear models. *J. Amer. Statist. Assoc.* **81**, 96-103.

Sain, S. R., Baggerly, K. A. and Scott, D. W. (1994). Cross-validation of multivariate densities. *J. Amer. Statist. Assoc.* **89**, 807-817.

Silverman, B. W. (1978). Density ratios, empirical likelihood and cot death. *Appl. Statist.* **27**, 26-33.

Silverman, B. W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *Ann. Statist.* **10**, 795-810.

Stone, C. (1994). The use of polynomial splines and their tensor products in multivariate function estimation (with discussion). *Ann. Statist.* **22**, 118-184.

Wahba, G. (1990). *Spline Models for Observational Data*, CBMS–NSF Regional Conference Series in Applied Mathematics, Vol. 59, Philadelphia: SIAM.

Wang, M.-C. (1989). A semiparametric model for randomly truncated data. *J. Amer. Statist. Assoc.* **84**, 742-748.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Johny Wiley.

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

Department of Statistics, Purdue University, West Lafayette, IN 47907, U.S.A.

E-mail: chong@stat.purdue.edu

(Received April 1996; accepted May 1997)