# Structural Multivariate Function Estimation: Some Automatic Density and Hazard Estimates

by

Chong Gu

Technical Report #93–28

Department of Statistics
Purdue University

June 1993

# Structural Multivariate Function Estimation: Some Automatic Density and Hazard Estimates

CHONG GU*

## 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 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, an 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, an exotic 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: ANOVA decomposition; Density estimation; Hazard estimation; Penalized likelihood; Performance-oriented iteration; 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".

*Chong Gu is Assistant Professor, Department of Statistics, Purdue University, West Lafayette, IN 47907. The author thanks Dr. Richard Moore for permission to use the ZVD treatment data and Dr. Mei-Cheng Wang for help in communication and in transmitting the data. This research was supported by grants from National Science Foundation and from Purdue Research Foundation.

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, penalized likelihood method pioneered by Good and Gaskins (1971) allows convenient structural model construction, and hence is rather powerful in handling multivariate problems.

Let $\eta$ be a function of interest, smooth models 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 $\int \ddot{\eta}^2$. When $\rho = 0$, $M_0 = J_\perp$ defines a parametric model. As $\rho$ increases, $M_\rho$ allows more and more flexible fits. To fit the model in $M_\rho$, one usually resort to the maximum likelihood method. The estimate in $\{\eta : J(\eta) \leq \rho\}$ 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 $\lambda$ is called the smoothing parameter, which controls the tradeoff between the goodness-of-fit and the smoothness of $\eta$.

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* $\eta$ *and a possibly unknown nuisance parameter* $\phi$*. Observing independent data* $(X_i, Y_i)$*,* $i = 1, \cdots, n$*, the method estimates* $\eta$ *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* $X_i$*,* $i = 1, \cdots, n$*, from a probability density* $f(x)$ *supported on a finite domain* $\mathcal{X}$*, the method estimates* $f$ *by* $e^\eta / \int_\mathcal{X} e^\eta$*, where* $\eta$ *minimizes*

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

*A side condition, say* $\int_\mathcal{X}\eta = 0$*, shall be imposed on* $\eta$ *for a 1-1 transform* $f \leftrightarrow e^\eta / \int_\mathcal{X} e^\eta$*.*

**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 $(Z_i, X_i, \delta_i, U_i)$, $i = 1,\cdots,n$, where $X = \min(T,C)$, $\delta = I_{[T\leq C]}$, and $Z < X$, the method estimates the log hazard $\eta$ via minimizing*

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

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 Good and Gaskins (1971), Leonard (1978), and Silverman (1982). The formulation (1.4) of Gu (1992b) bears influence from 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 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 function of $x$ is defined as $\eta(x) = g_{\emptyset} + g_t + g_u + g_{t,u}$, where $g_{\emptyset}$ 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 via a certain structural construction of $J$ known as tensor-product spline technique, of which an exposition can be found in Gu and Wahba (1992). 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 $\lambda$ but a completely specified $J$ have been developed in previous work (Gu 1993a, b). With an ANOVA decomposition built in, say $\eta = \sum_\beta g_\beta$ where $\beta$ is a generic index, however, $J$ is usually of the form $\sum_\beta \theta_\beta^{-1} J_\beta(g_\beta)$, where $J(g_\beta)$ measure the roughness of function components $g_\beta$, and the weights $\theta_\beta$, an extra set of smoothing parameters, should naturally also be selected adaptively. After a brief review of existing results, I shall discuss a feasible automatic multiple smoothing parameter algorithm for calculating density and hazard estimates of (1.3) and (1.4), and with the help of the algorithm, I shall illustrate some new options of nonparametric multivariate data analysis made possible by the technique.

## 2 Formulation and Preliminaries

In this section, I shall 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.

Of 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 $\mathcal{H} \subseteq \{\eta : J(\eta) < \infty\}$. $J(\eta)$ 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 $\eta$, 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 $\lambda$.

I shall 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 I shall 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 $J = \int \ddot{\eta}^2$ with $J_\perp = \{1, (\cdot - .5)\}$. The function space $\{g : \int \ddot{g}^2 < \infty\}$ 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 $\mathcal{H}_s = \{g : \int \ddot{g}^2 < \infty, \int g = \int \dot{g} = 0\}$ has an RK $R_s(t, s) = k_2(t)k_2(s) - k_4(|t - s|)$ dual to the norm $J(g) = \int \ddot{g}^2$, where $k_\nu = B_\nu/\nu!$ and $B_\nu$ are the $\nu$th Bernoulli polynomials (cf. Craven and Wahba 1979). A univariate ANOVA decomposition is in place where $\mathcal{H}_c$ carries constant and $\mathcal{H}_\pi \oplus \mathcal{H}_s$ carries the "treatment effect" satisfying the side condition $\int g = 0$. 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 simply 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 $J_\beta$ 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 $J_\perp$. $\mathcal{H}_J = \oplus_{\theta_\beta \in (0,\infty)} \mathcal{H}_\beta$ and $R_J = \sum_{\theta_\beta < \infty} \theta_\beta R_\beta$. $\mathcal{H}_{c,c}$, $\mathcal{H}_{c,\pi}$, $\mathcal{H}_{\pi,c}$, 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 1-1 logistic density transform $f \leftrightarrow e^\eta / \int e^\eta$. Formulas of $J_\beta$ in this construction can be found in Gu and Wahba (1992) but are not needed for computation.

When $L(\eta)$ depends on $\eta$ 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 $\eta$.

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 $\eta = \sum_{i=1}^n c_i \xi_i + \sum_{\nu=1}^M d_\nu \phi_\nu$. Fixing smoothing parameters, $\hat{\eta}_n$ can be calculated by minimizing

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

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 $(i,\nu)$th entry $\phi_\nu(X_i)$. Let $\mu_\eta(h) = \int h e^\eta / \int e^\eta$ 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 $V_{\phi,\eta} = V_{\tilde{\eta}}(\boldsymbol{\phi}, \tilde{\eta})$; details can be found in Gu (1993a).

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 $\lambda$ and $\eta$ was developed in Gu (1993a), where the performance is measured by a computable proxy (cf. 3.1) of the symmetrized Kullback-Leibler between the true density and the estimate, $\mathrm{SKL}(\eta, \eta_0) = \mu_{\eta_0}(\eta_0 - \eta) + \mu_\eta(\eta - \eta_0)$. The arguments and formulas in Gu (1993a, Section 3) remain valid when $\theta_\beta$ hidden in $R_J$ are also to be updated, but the single smoothing parameter algorithm of Gu (1993a, Section 4) is no longer sufficient. We shall discuss a multiple smoothing parameter algorithm in the next section.

Parallel results hold for hazard estimation. Let $N = \sum_{i^*=1}^n \delta_{i^*}$ and $T_i$, $i = 1, \cdots, N$, 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 $\mathcal{H}_n = J_\perp \cup \{R_J((T_i, U_i), \cdot)\}$ share the same asymptotic convergence rates as $\hat{\eta}$ in $\mathcal{H}$ (cf. Gu 1992b). $\hat{\eta}_n = \sum_{i=1}^N c_i R_J((T_i, U_i), \cdot) + \sum_{\nu=1}^M d_\nu \phi_\nu(\cdot) = \boldsymbol{\xi}^T\boldsymbol{c} + \boldsymbol{\phi}^T\boldsymbol{d}$ can be computed via minimizing

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

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 $(i,\nu)$th entry $\phi_\nu(T_i, U_i)$, $Y_{i^*}(t) = I_{[X_{i^*} \geq t > Z_{i^*}]}$ is the at-risk process of the $i^*$th observation, $\boldsymbol{\xi}_{i^*}$ is $N \times 1$ with

$i$th entry $\xi_i(t, U_{i^*})$, and $\boldsymbol{\phi}_{i^*}$ is $M \times 1$ with $\nu$th entry $\phi_\nu(t, U_{i^*})$. The one-step Newton update for minimizing (1.4) again satisfies (2.2) but with the entries modified as entailed from the modified definitions of $\mu_\eta(h) = (1/n)\sum_{i^*=1}^{n} \int_{\mathcal{T}} h_{i^*} Y_{i^*} e^{\eta_{i^*}}$ and $V_\eta(h, g) = \mu_\eta(hg)$, where $h_{i^*}(t) = h(t, U_{i^*})$ and $\eta_{i^*}(t) = \eta(t, U_{i^*})$. As a proxy of $\mathrm{SKL}(\eta, \eta_0) = \int_{\mathcal{U}} \int_{\mathcal{T}} (e^\eta - e^{\eta_0})(\eta - \eta_0)\tilde{S}m$ where $\tilde{S}(t, u) = P(X \geq t > Z | U = u)$ is the at-risk probability and $m(u)$ is the density of $U$, (3.1) holds verbatim for hazard estimation up to the entries appearing in (2.2); an argument can be found in Gu (1993b) for a singleton $\mathcal{U}$, which extends readily to the general setup.

# 3 Algorithm

Define $H = V_{\xi,\xi} + \lambda Q$, $E = V_{\phi,\phi} - V_{\phi,\xi} H^{-1} V_{\xi,\phi}$, $\boldsymbol{u}_\xi = Q\mathbf{1}/n - \boldsymbol{\mu}_\xi + V_{\xi,\eta}$, $\boldsymbol{u}_\phi = S^T\mathbf{1}/n - \boldsymbol{\mu}_\phi + V_{\phi,\eta}$, $\boldsymbol{u}_{\phi|\xi} = \boldsymbol{u}_\phi - V_{\phi,\xi} H^{-1} \boldsymbol{u}_\xi$, $\boldsymbol{v}_\xi = V_{\xi,\eta} - \boldsymbol{\mu}_\xi$, $\boldsymbol{v}_\phi = V_{\phi,\eta} - \boldsymbol{\mu}_\phi$, and $\boldsymbol{v}_{\phi|\xi} = \boldsymbol{v}_\phi - V_{\phi,\xi} H^{-1} \boldsymbol{v}_\xi$. Solving (2.2) one gets $\boldsymbol{d} = E^{-1} \boldsymbol{u}_{\phi|\xi}$ and $\boldsymbol{c} = H^{-1}(\boldsymbol{u}_\xi - V_{\xi,\phi}\boldsymbol{d})$, which are dependent on the smoothing parameters $\lambda$ and $\theta_\beta$ hidden in $R_J$. It is shown in Gu (1993a, Section 3) that

$$\begin{aligned} \hat{L}_{\tilde{\eta}}(\eta, \eta_0) &= \frac{\mathrm{trace}(QH^{-1}Q)}{n(n-1)} - \frac{(Q\mathbf{1}/n)^T H^{-1} (Q\mathbf{1}/n)}{n-1} - \frac{\boldsymbol{u}_\xi^T H^{-1} \boldsymbol{u}_\xi + \boldsymbol{u}_{\phi|\xi}^T E^{-1} \boldsymbol{u}_{\phi|\xi}}{2} \\ &\quad - \frac{\lambda\,(\boldsymbol{u}_\xi - V_{\xi,\phi} E^{-1} \boldsymbol{u}_{\phi|\xi})^T H^{-1} Q H^{-1} (\boldsymbol{u}_\xi - V_{\xi,\phi} E^{-1} \boldsymbol{u}_{\phi|\xi})}{2}, \end{aligned} \tag{3.1}$$

is a proxy of $\mathrm{SKL}(\eta, \eta_0)$, where $\eta = \boldsymbol{\xi}^T \boldsymbol{c} + \boldsymbol{\phi}^T \boldsymbol{d}$ is dependent on $\lambda$ and $\theta_\beta$ through $\boldsymbol{\xi}$, $\boldsymbol{c}$, and $\boldsymbol{d}$. A performance-oriented iteration may be conducted by selecting $\lambda$ and $\theta_\beta$ to minimize (3.1) in each iteration.

The algorithm I will be using consists of an initialization step and an updating step. For the initial value of $\theta_\beta$, say $\theta_1$, I first set $\theta_1 = 1$ and $\theta_\gamma = 0$, $\gamma \neq 1$, use the single smoothing parameter algorithm of Gu (1993a, Section 4) to obtain an automatic $\lambda$, say $\lambda_1$, with $R_1$ being the only penalized term, and then set $\theta_1 = 1/\lambda_1$. After separate calculations of initial $\theta_\beta$ in this manner, I put all penalized terms back together and employ the fixed $\theta_\beta$ algorithm once more to set up for the updating step. Such a procedure is invariant to individual scalings of $R_\beta$, which are usually arbitrary and not comparable to each other, and if the penalized terms contribute somewhat "independently" to the estimate, the relative weights should not be too far from the "optimal" ones.

In each iteration of the updating step, I first fix $\lambda$ and $\tilde{\eta}$ and update $\theta_\beta$ one at a time through the list, then I invoke the fixed $\theta_\beta$ algorithm to calculate $\lambda$ and $\tilde{\eta}$ for the next iteration. When updating a certain $\theta_\beta$, say $\theta_1$, I fix other $\theta_\beta$ at their latest values, evaluate (3.1) at three different values of $\theta_1$: the current value, say $\tilde{\theta}_1$, and two adjacent values $\tilde{\theta}_1 10^{\pm .1}$; I then fit a quadratic in $\log_{10}\theta_1$ through the three points, determine the minimum of the quadratic on $[\log_{10}\tilde{\theta}_1 - .5, \log_{10}\tilde{\theta}_1 + .5]$, and evaluate (3.1) once more at the minimum; the smallest of the four evaluations of (3.1) gives the new $\theta_1$. The order in which $\theta_\beta$ is updated could be arbitrary, but for definiteness I 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 $\theta_\beta$ matter so a standardization procedure should follow the $\theta_\beta$ updating, for which I choose to set the trace of $Q$ to one. The algorithm is clearly invariant to the scaling and indexing of $R_\beta$.

I choose such a simple coordinate-wise updating procedure out of the following considerations. First, the derivatives of (3.1) are beyond reach so the Newton method is not feasible. Second, (3.1) will change from iteration to iteration and so will the minimizing $\theta_\beta$, hence it would be unwise to invest too much to try to minimize (3.1) in one iteration; this rules out the usual quasi-Newton approach. Now if $g_\beta$ contribute "independently" to $\eta$, one may expect (3.1) not to have too much curvature in the neighborhood of the minimum, as suggested by the behavior of a similar score in the regression setup (cf. Gu, Bates, Chen and Wahba 1989), so a coordinate-wise updating can be reasonably efficient.

The algorithm is implemented in portable RATFOR code soon to be released to public archives. Its performance will be discussed along with the examples in sections to follow. Numerical details are not of prime interest here and are omitted.

## 4 Density Estimation Examples

I shall first analyze the blood fat concentration data listed in Scott (1992, Appendix B.3). Concentrations of plasma cholesterol and plasma triglycerides (mg/dl) in 371 male patients were evaluated for chest pain. The patients were classified into two groups, 51 patients with no evidence of heart disease, and 320 patients with narrowing of the arteries. Of interest are the estimation and comparison of the bivariate distributions of cholesterol and triglycerides for the two groups.

I transformed both variables by $\log_{10}$ to make the data more evenly scattered, and employed the algorithm of Section 3 to estimate the density for the disease group. Let $T = \log_{10}(\text{cholesterol})$ and

$U = \log_{10}(\text{triglycerides})$. Two outliers were removed (see below for discussion) from the disease group and a domain $\mathcal{X} = \mathcal{T} \times \mathcal{U} = [2, 2.65] \times [1.5, 2.9]$ was chosen to cover the remaining 369 observations. The support of the density is often unknown in practice whereas a finite domain has to be specified for the method to work, but the resulting estimate can (and should) always be considered as that of the conditional distribution of $X|(X \in \mathcal{X})$. $\mathcal{X}$ was mapped onto $[0,1]^2$ for calculation using the tensor product spline construction of Section 2. On the translated domain $[0,1]^2$, I 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 $g_t = d_1\phi_1(x) + \sum_{i=1}^{n} \theta_{s,c}R_{s,c}(X_i, x)c_i = d_1(t-.5) + \sum_{i=1}^{n} \theta_{s,c}R_s(T_i, t)c_i$, $g_u = d_2\phi_2(x) + \sum_{i=1}^{n} \theta_{c,s}R_{c,s}(X_i, x)c_i = d_2(u-.5) + \sum_{i=1}^{n} \theta_{c,s}R_s(U_i, u)c_i$, and $g_{t,u} = d_3\phi_3(x) + \sum_{i=1}^{n}(\theta_{s,\pi}R_{s,\pi}(X_i, x) + \theta_{\pi,s}R_{\pi,s}(X_i, x) + \theta_{s,s}R_{s,s}(X_i, x))c_i = d_3(t-.5)(u-.5) + \sum_{i=1}^{n}(\theta_{s,\pi}R_s(T_i, t)(U_i - .5)(u-.5) + \theta_{\pi,s}(T_i - .5)(t-.5)R_s(U_i, u) + \theta_{s,s}R_s(T_i, t)R_s(U_i, u))c_i$. All the integrals appearing in the terms of (2.2) were calculated by summation over an equally spaced $50 \times 50$ grid on $\mathcal{X}$; see Gu (1993a) for discussion concerning the choice of quadrature in this setting.

The estimated density for the disease group is contoured in frame (a) of Figure 4.1 as solid lines, where the 318 observations used in the estimation are superimposed as circles, the 2 outliers as stars, and the integration grid as dots. The estimated density for the healthy group is similarly contoured in frame (b) of Figure 4.1 and observations and integration grid superimposed. Frame (c) overlays the two estimated densities and frame (d) contours the likelihood ratio of the disease group density over the healthy group density. To assess the strength of the dependence of the two variables, I calculated the information proper $\text{Inf}(T \perp U) = E_{f_{TU}} \log(f_{TU}/f_T f_U)$ as defined in Whittaker (1990, Chapter 4), otherwise known as the relative entropy (cf. Joe 1989), where $f_{TU}$ is the estimated joint density and $f_T$ and $f_U$ are the corresponding marginals. The information proper were .04118 and .04285 for the disease and healthy groups respectively, and the equivalent normal correlations $\rho = \{1 - \exp(-2\text{Inf})\}^{1/2}$ are .281 and .287.

For the second example 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"

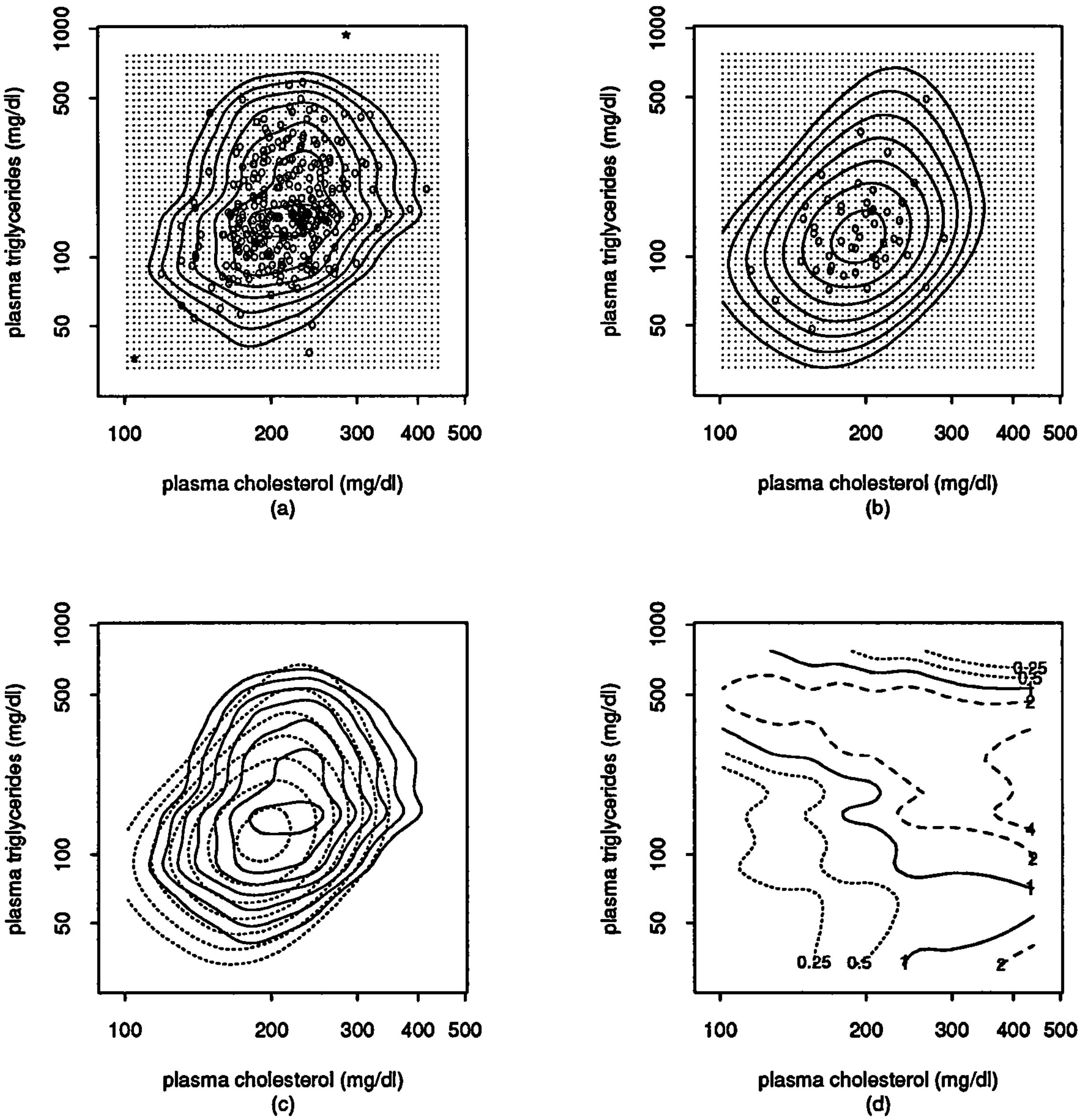

Figure 4.1: Blood Fat Concentration Data. (a) Disease group density with data. (b) Healthy group density with data. (c) Overlay of densities of (a) and (b). (d) Likelihood ratio of (a) over (b).

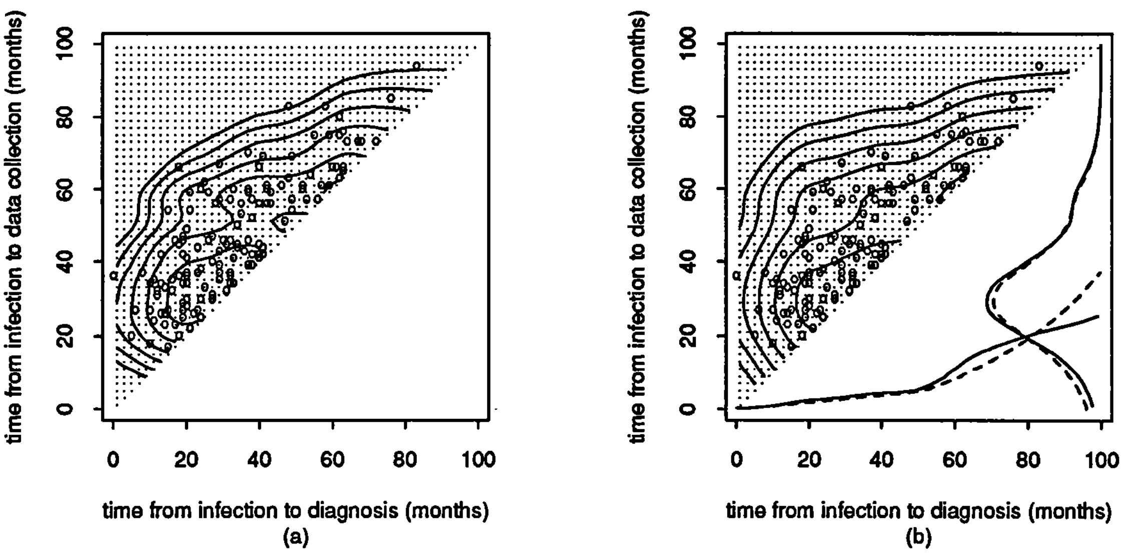

Figure 4.2: Blood-Transfusion Data of "Elderly Patients". (a) Density estimate without pre-truncation independence. (b) Density estimates with pre-truncation independence.

of age 60 or above. Clearly only data with $T \leq U$ 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 (1993c) by a penalized conditional likelihood approach. Using penalized full likelihood with multiple smoothing parameters, one may check the assumption of pre-truncation independence by allowing interaction in the estimation and assessing its strength. Under the assumption of pre-truncation independence, penalized full likelihood with multiple smoothing parameters estimates the distributions of $T$ and $U$ simultaneously. For illustration I shall conduct these analyses for the "elderly patients" and compare the results with those in Gu (1993c).

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 $\mathcal{X} = [0, 100]^2 \cap \{t \leq u\}$. I mapped $[0, 100]^2$ onto $[0, 1]^2$ and used the same $\phi_\nu$'s and RK components as those used in the previous example. The estimated density without pre-truncation independence is contoured in frame (a) of Figure 4.2 with the data superimposed as circles and the integration grid as dots, where the grid points on the diagonal carried half weights as compared to the others in calculating the integrals. The estimate with pre-truncation independence was calculated by setting $d_3 = \theta_{s,\pi} = \theta_{\pi,s} = \theta_{s,s} = 0$ and is contoured in frame (b) of Figure 4.2 where the data and the integration grid are superimposed. The distributions on the margins are superimposed in the blank space on their corresponding axes, where the solid lines are estimates based on full likelihood calculated here and the dashed lines are estimates based on conditional likelihoods quoted from Gu (1993c). The heights of the solid and dashed lines are adjusted so the areas under them are the same. It is clear that the two estimates agree well in the light truncation area but depart a bit in the heavy truncation area. To assess the feasibility of pre-truncation independence, I calculated the log likelihood ratio $\sum_{i=1}^{141} \log(f_1(X_i)/f_2(X_i)) = 7.243$ where $f_1$ and $f_2$ are the estimates without and with pre-truncation independence. This gives a "$\chi^2$-statistic" about 14.5 but one needs an appropriate degrees of freedom for its calibration. A closer look at the components of the estimate revealed that $d_3\phi_3$ dominated the interaction term in $\log f_1$, so one degree of freedom appears reasonable for the calibration of this particular "$\chi^2$-statistic" resulting from nonparametric estimation with an associated $p$-value of .000, and in turn the pre-truncation independence does not seem to be too good an assumption for the observed data. An alternative check on the interaction is to set $d_3 = 0$ and see if the penalized terms are "willing" to reproduce the effect of $d_3\phi_3$ in $f_1$. Denoting this estimate as $f_3$, we have a log likelihood ratio $\sum_{i=1}^{141} \log(f_1(X_i)/f_3(X_i)) = .179$, so the same effect was indeed reproduced. The Kullback-Leibler information are $E_{f_1} \log(f_1/f_2) = .02578$ and $E_{f_2} \log(f_2/f_1) = .03180$ yielding $\text{SKL}(f_1, f_2) = .05758$, $E_{f_3} \log(f_3/f_2) = .01717$ and $E_{f_2} \log(f_2/f_3) = .01975$ yielding $\text{SKL}(f_3, f_2) = .03692$, and $E_{f_1} \log(f_1/f_3) = .00211$ and $E_{f_3} \log(f_3/f_1) = .00224$ yielding $\text{SKL}(f_1, f_3) = .00435$.

Finally comes some discussion on the performance of the algorithm. The performance-oriented iteration implemented in the algorithm operates on mathematically different performance proxies at different $\tilde{\eta}$, so convergence is not guaranteed. Nevertheless, divergence rarely occurs in my experiments with a single smoothing parameter (cf. Gu 1993a, b, c). Multiple smoothing parameters introduce greater flexibility, however, and one may expect a bit more difficulties. For the blood-transfusion data, the algorithm converged in all cases without incidence. For the blood fat concentration data, the algorithm converged for the healthy group and for the disease group minus 2 outliers to give the estimates presented, but it did not converge when the outliers were kept in the disease group data: The algorithm iterated to a certain $\theta$ combination on which the fixed $\theta$ iteration asked for interpolation. Similar phenomena also occurred in runs with untransformed or partly transformed data with apparent outliers. Because a key step in calculating performance proxies is the estimation of terms of $\mu_{\eta_0}(\eta)$ via the sample mean or cross-validated version of it (cf. Gu 1993a, Section 3), it appears reasonable for the procedure to be somewhat sensitive to outliers, but how exactly this may cause trouble to $\theta$ updating remains unclear. Numerically the algorithm is quite demanding: A run for the $n = 318$ blood fat concentration example with 5 $\theta$'s took about 300 cpu minutes out of an IBM-RS6000 and a run for the $n = 141$ blood-transfusion example with 5 $\theta$'s took about 30 cpu minutes.

# 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 censerings. There is no truncation in the data, i.e., $Z_i \equiv 0$. Due to the nonsignificance in the analyses by Miller and Halpern (1982) and by others, and also due to the missing values, I shall discard the tissue type mismatch score in the following analysis.

Let $T$ be time after transplant and $U$ be age at transplant. I transformed the time axis by $t^* = t^{1/2}$ to make the survival/censering times more evenly scattered, and then estimated hazard 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 $e^{\eta(t^*,u)}(dt^*/dt) = e^{\eta(t^{1/2},u)}/(2t^{1/2})$. I mapped $\mathcal{T}^* \times \mathcal{U}$ 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 solid lines in frame (a) of Figure 5.1 and that without interaction in frame (b), where the contour labels are multiplied by 100 and data are superimposed as circles (deceased) or crosses (censered). To assess the plausibility of hazard proportionality, I calculated the log likelihood ratio $\sum_{i=1}^{184}\{\delta_i(\eta_1 - \eta_2)(X_i, U_i) - \int_0^{X_i}(e^{\eta_1(t,U_i)} - e^{\eta_2(t,U_i)})dt\} = 3.376$ (cf. 1.4) which yields a "$\chi^2$-statistic" 6.65, where $\eta_1$ is the fit of frame (a) and $\eta_2$ is the proportional hazard fit of frame (b). This time the penalized terms did contribute to the interaction so one degree of freedom is only a lower bound for the calibration of this particular "$\chi^2$-statistic". The $p$-values associated with $\chi^2 = 6.65$ are .010, .036, and .084 for $d.f. = 1, 2, 3$, so

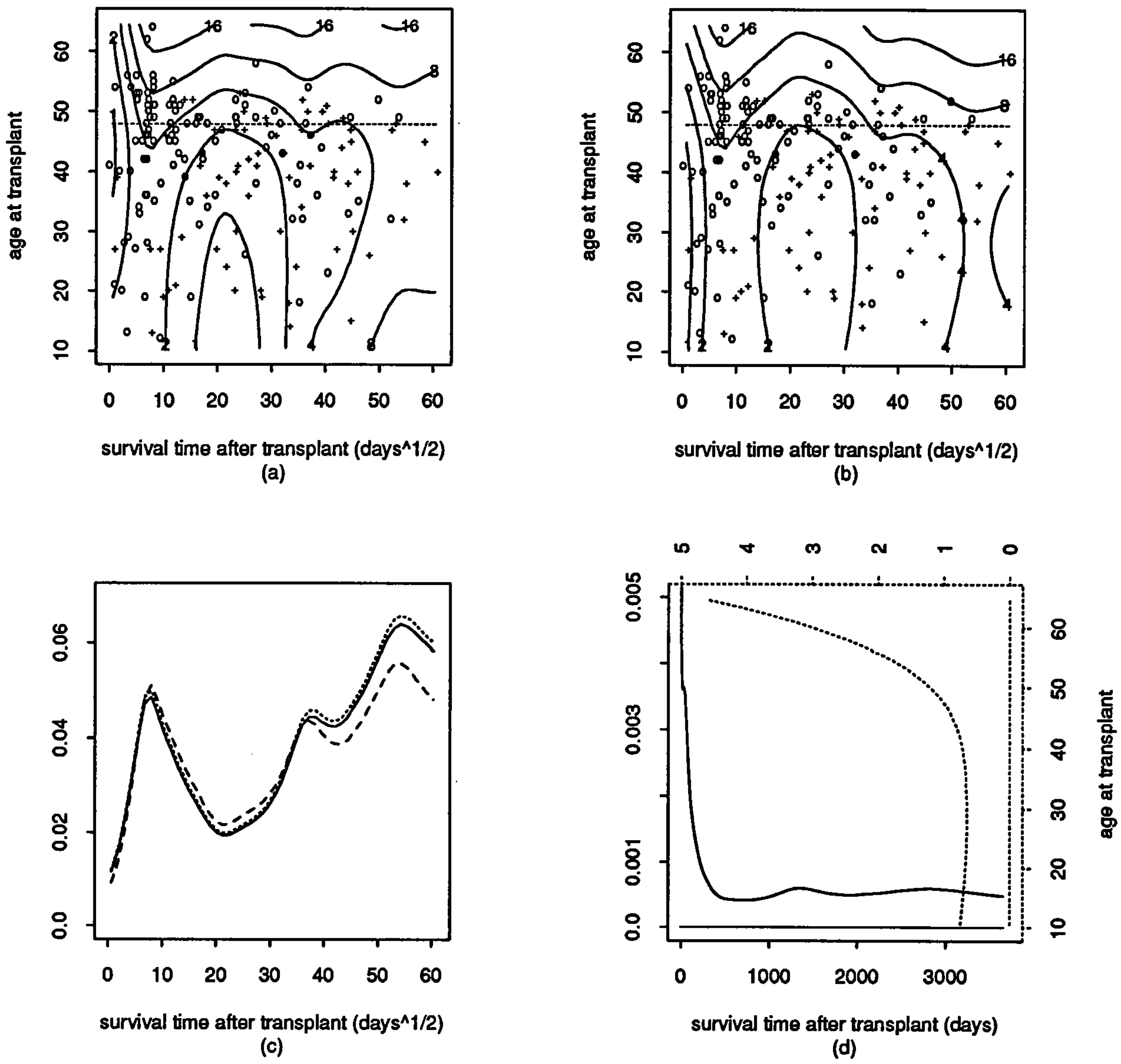

Figure 5.1: Stanford Heart Transplant Data. (a) $e^{\eta(t^*,u)}$ with interaction. (b) $e^{\eta(t^*,u)}$ without interaction (proportional hazard $e^{g_\emptyset+g_{t^*}}e^{g_u}$). (c) "Base hazard" $e^{g_\emptyset+g_{t^*}(t^*)}$ of (b) and slices of (a) and (b). (d) "Base hazard" on the original time axis $e^{g_\emptyset+g_{t^*}(t^{1/2})}/(2t^{1/2})$ and age multiplier $e^{g_u}$ of (b). Circles in (a) and (b) are observed deaths and crosses censorings.

the interaction seems at most marginally significant, and in turn hazard proportionality appears plausible. Plotted in frame (c) of Figure 5.1 are the "base hazard" $e^{g_\emptyset+g_{t^*}}$ of the proportional hazard model as the solid line, a slice from frame (b) on the dotted line as the dotted line, and a slice from frame (a) on the same dotted line as the dashed line; data are relatively dense near the sliced line in $\mathcal{T}^* \times \mathcal{U}$ and the two slices agree well up to about $t^* = 35$, beyond which only a few failures occurred near the sliced line so information from data is very limited. The "base hazard" on the original time axis and the age multiplier $e^{g_u}$ in the proportional hazard model share frame (d) 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.

For the second example I shall try to analyze the Zidovudine (ZVD) treatment data described by Wang, Brookmeyer, and Jewell (1993). A total of 500 patients with a prior diagnosis of AIDS were recruited for a two-year observational study of ZVD treatment sponsored by the Burroughs Wellcome Company. Available are time from AIDS diagnosis to enrollment in the study, time from enrollment to end of follow-up, both in days, and death/censoring indicator. There were 206 recorded deaths and 294 censorings. One may set the time origin at AIDS diagnosis (incident model) or at initiation of treatment (prevalent model); see Wang, Brookmeyer, and Jewell (1993) for relevant discussion. When the time origin is set at AIDS diagnosis, the life time data are truncated at enrollment time. I shall use $U = \log_{10}(\text{time from AIDS diagnosis to enrollment}+10)$ as a covariate, and estimate hazards in both the prevalent model and the incident model. The log transform on covariate makes the data more evenly scattered.

Let $T_1$ be follow-up time and $T_2$ be time since AIDS diagnosis. Data are on $(t_1, u) \in [0, 735] \times [1, 3.2]$ and $(t_2, u) \in [0, 1570] \times [1, 3.2]$, respectively, on the two time axes. The domains were mapped onto $[0,1]^2$ and the same tensor product spline construction was used to calculate estimates. Automatic hazard estimates are contoured in Figure 5.2 with contour labels multiplied by 1000, where the two frames in left are for the prevalent model and those in right for the incident model. The frames on top carry interaction terms and the frames on bottom assume hazard proportionality. Observed deaths are superimposed as circles and censorings as crosses. The truncation line is also superimposed as the dotted lines for the incident model. An assumption for the theory and the

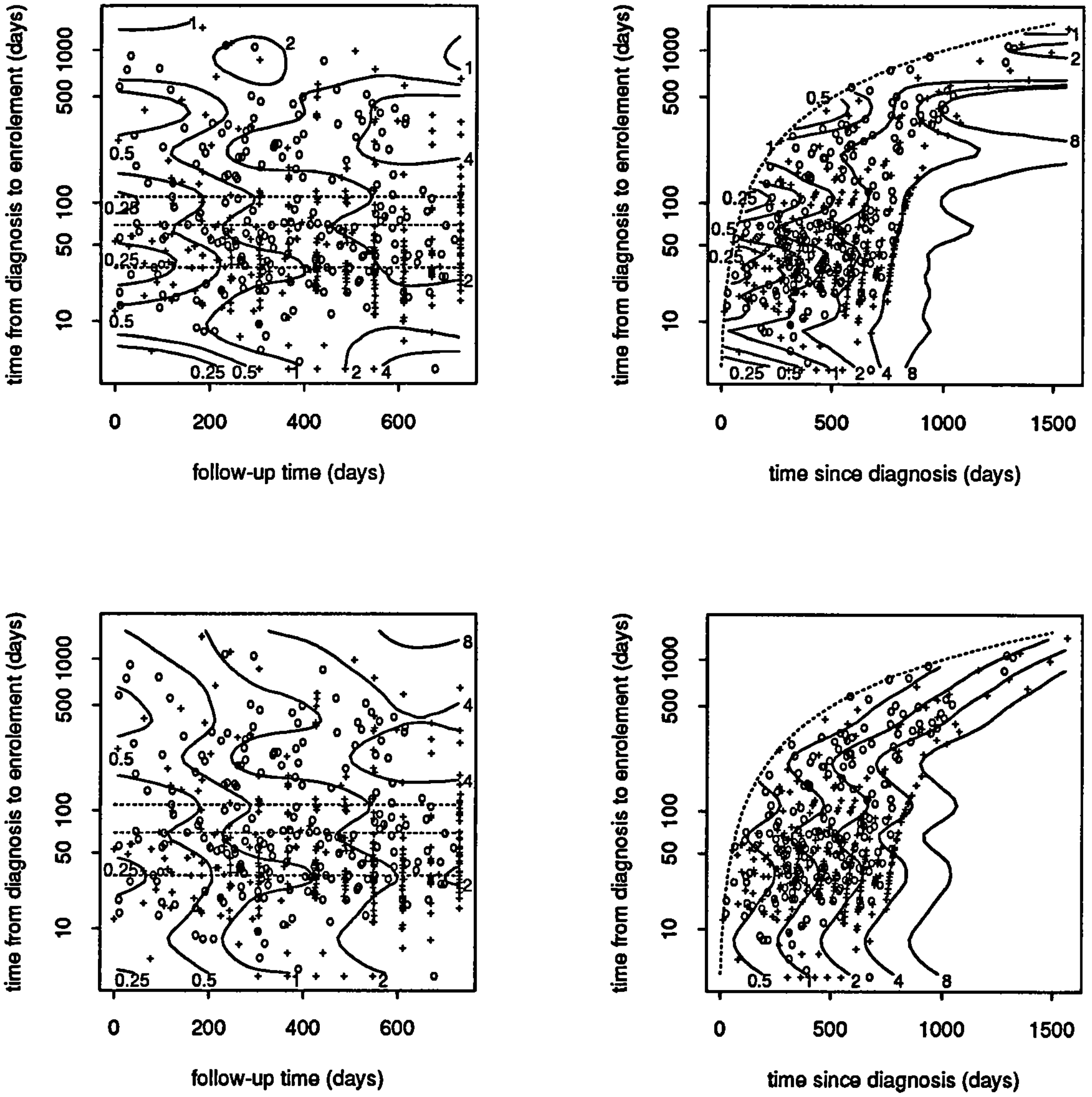

Figure 5.2: Zidovudine AIDS Treatment Data. Left: estimates in prevalent model; right: estimates in incident model. Top: estimates with interaction; bottom: proportional hazard estimates. Circles are observed deaths and crosses censorings.

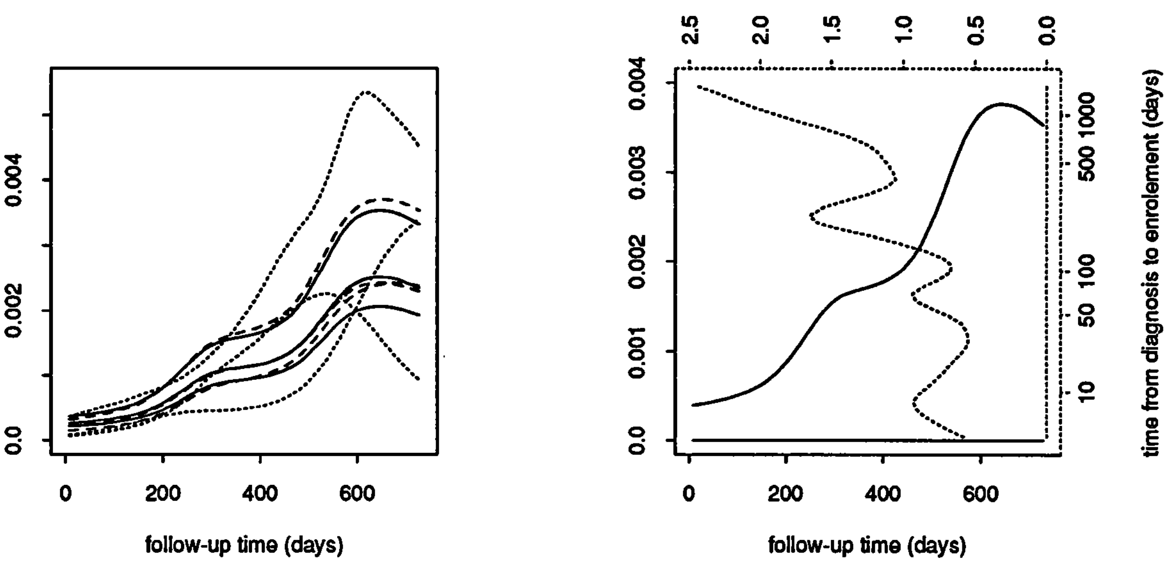

Figure 5.3: Zidovudine AIDS Treatment Data, Prevalent Model. Left: slices of estimates; solid lines are from proportional hazard estimate, dashed lines from estimate with interaction, and dotted lines from discretized data. Right: base-hazard $e^{g_\emptyset+g_t}$ and covariate effect $e^{g_u}$ in the proportional hazard estimate.

performance estimate to work properly is that the truncation times (and life and censoring times) be independent of the covariate (cf. Gu 1992b, 1993b), which does not hold for the incident model in this example. This may explain the apparent undersmoothing in the upper-right frame of Figure 5.2, yet the proportional hazard estimate in the lower-right frame appears reasonable.

We shall take a closer look at the prevalent model. The log likelihood ratio of the estimate with interaction over the proportional hazard estimate is 7.292 yielding a "$\chi^2$-statistic" 14.58, which has a $p$-value of .010 with a $d.f.$ = 4.75. We are in a grey area with little guidance, and the "$\chi^2$-statistic" does not seem to be of much use here. The left frame of Figure 5.3 compares the two estimates on three slices marked as dotted lines in the left frames of Figure 5.2, where the solid lines are from the proportional hazard estimate and the dashed lines from the estimate with interaction; the estimates seem to agree well in data-dense area. The base-hazard and the covariate effect in the proportional hazard estimate are plotted in the right frame of Figure 5.3 in the same manner as in frame (d) of Figure 5.1. We see that the hazard increases as time goes on, but the up-and-downs in the covariate effect appears puzzling. To check whether these are "real" or just some artifacts due to the estimation procedure, I selected three subsets of data with time from diagnosis to enrollment in the ranges of $[30, 40]$, $[60, 80]$, and $[95, 130]$ surrounding the three slices marked in the left frames of Figure 5.2, and estimated log hazards for the three groups using cubic splines as in Gu (1993b); the sample sizes and observed deaths are 79/28, 50/27, and 32/8 for the three groups in order. To make things somewhat comparable, I turned off the automatic smoothing parameter selection, and instead chose to match the integrated second derivative of the estimated log hazards to that of $g_t$ in the proportional hazard estimate. The three estimates are superimposed in the left frame of Figure 5.3 as dotted lines, where referring to the upper end of the time axis, the upper, middle, and lower curves correspond to covariate ranges $[60, 80]$, $[95, 130]$, and $[30, 40]$. A bump around 70 on the covariate axis is clearly there. We also see that separate estimates go their own ways, especially on the upper end of time axis where information from data is scarce, while smoothing on the covariate axis reconciles them and hopefully reduces noise in turn. The up-and-downs in the covariate effect may or may not be "real", but they appear to faithfully describe features of the data.

All iterations for the hazard examples 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 the Stanford heart transplant data, the fit with interaction took 50 cpu minutes on an IBM-RS6000 and the proportional hazard fit took 29. For the ZVD treatment data, the two fits in the top row of Figure 5.2 took 974 cpu minutes in total and those in the bottom took 375.

## 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 estimations soon to be released to public domain archives statlib and netlib, as sequel to a collection of routines for smoothing spline regression archived in RKPACK.

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 estimator respects qualitatively different axes and the automatic selection of smoothing parameters makes the estimation invariant of 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 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 a certain estimated performance of the fit. Kooperberg, Stone and Truong (1993) 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 represent a modest step forward towards nonparametric estimation of graphical models (cf. Whittaker 1990); related work on conditional density estimation is under way. For hazard estimation, a further topic is the incorporation of time dependent covariate, on which a theory has yet to be developed.

If observed $(T, U)$ center around some monotone curve in $\mathcal{T} \times \mathcal{U}$, the estimated terms in an ANOVA decomposition may suffer identifiability problem; this is 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 practical strengths of estimated ANOVA terms. Some diagnostic tools for regression can be found in Gu (1992a). Those for density and hazard estimations are yet to be developed.

## References

Aronszajn, N. (1950), "Theory of reproducing kernels," *Transactions of American Mathematical Society*, 68, 337 – 404.

Cox, D. R. (1972), "Regression models and life tables" (with discussion), *Journal of the Royal Statistical Society 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," *Numerische Mathematik*, 31, 377 – 403.

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," *Journal of the American Statistical Association*, 87, 942 – 951.

Gu, C. (1992a), "Diagnostics for nonparametric regression models with additive terms," *Journal of the American Statistical Association*, 87, 1051 – 1058.

—— (1992b), "Penalized likelihood hazard estimation: A general asymptotic theory," Technical Report 91-58 (Rev.), Purdue University, Dept. of Statistics.

—— (1993a), "Smoothing spline density estimation: A dimensionless automatic algorithm," *Journal of the American Statistical Association*, 88, 495 – 504.

—— (1993b), "Penalized likelihood hazard estimation: Algorithm and examples," In *Statistical Decision Theory and Related Topics V*, ed. S. S. Gupta and J. Berger, Springer-Verlag, pp 000 – 000.

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

Gu, C., Bates, D., Chen, Z., and Wahba, G. (1989), "The computation of GCV functions through Householder tridiagonalization with applications to the fitting of interaction spline models," *SIAM Journal on Matrix Analysis and Applications*, 10, 457 – 480.

Gu, C. and Qiu, C. (1993), "Smoothing spline density estimation: Theory," *The Annals of Statistics*, 21, 217 – 234.

Gu, C. and Wahba, G. (1992), "Smoothing splines and analysis of variance in function spaces," Technical Report 91-29 (Rev.), Purdue University, Dept. of Statistics.

Joe, H. (1989), "Relative entropy measures of multivariate dependence," *Journal of the American Statistical Association*, 84, 157 – 164.

Kooperberg, C., Stone, C. J., and Truong, Y. K. (1993), "Hazard estimation," Technical Report 389, University of California–Berkeley, Dept. of Statistics.

Leonard, T. (1978), "Density estimation, stochastic processes and prior information" (with discussion), *Journal of the Royal Statistical Society 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 Journal on Scientific and Statistical Computing*, 9, 363 – 379.

—— (1988b), "Nonparametric estimation of relative risk using splines and cross-validation," *SIAM Journal on Scientific and Statistical Computing*, 9, 531 – 542.

Scott, D. W. (1992), *Multivariate Density Estimation: Theory, Practice and Visualization*, New York: Wiley.

Silverman, B. W. (1978), "Density ratios, empirical likelihood and cot death," *Applied Statistics*, 27, 26 – 33.

—— (1982), "On the estimation of a probability density function by the maximum penalized likelihood method," *The Annals of Statistics*, 10, 795 – 810.

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," *Journal of the American Statistical Association*, 84, 742 – 748.

Wang, M.-C., Brookmeyer, R., and Jewell, N. P. (1993), "Statistical models for prevalent cohort data," *Biometrics*, 49, 1 – 11.

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

Zucker, D. M. and Karr, A. F. (1990), "Nonparametric survival analysis with time-dependent covariate effects: A penalized partial likelihood approach," *The Annals of Statistics*, 18, 329 – 353.