Smoothing Spline Density Estimation: Conditional Distribution

by

Chong Gu
Purdue University

Technical Report #93–30

Department of Statistics
Purdue University

July 1993

△

SMOOTHING SPLINE DENSITY ESTIMATION: CONDITIONAL DISTRIBUTION

Chong Gu

Purdue University

Abstract

This article extends recent developments in penalized likelihood probability density estimation to the estimation of conditional densities. Positivity and unity constraints for a probability density is enforced through a one-to-one logistic conditional density transform made possible by term trimming in an ANOVA decomposition of multivariate functions. An asymptotic theory is sketched and the computation with automatic multiple smoothing parameters is noted. Examples are presented to illustrate possible applications of the technique. The development constitutes a viable approach to nonparametric estimation of graphical chain models, possibly with a mixture of continuous and categorical variables.

Key words and phrases: ANOVA decomposition, conditional distribution, density estimation, penalized likelihood, rate of convergence, regression, smoothing parameter.

1 Introduction

Let (X_i, Y_i), $i = 1, \cdots, n$, be independent observations from a probability density $f(x, y)$ on a product domain $\mathcal{X} \times \mathcal{Y}$. Of interest is the estimation of the conditional density $f(y|x) = f(x, y) / \int_{\mathcal{Y}} f(x, y)$ of Y given X, without assuming rigid constraints in the form of parametric models for $f(x, y)$ or $f(y|x)$. To achieve noise reduction in estimation, however, certain soft constraints on $f(x, y)$ or $f(y|x)$ are necessary. The method under study is the penalized likelihood method pioneered by Good and Gaskins (1971). The formulation follows that of Gu and Qiu (1993), which evolved from the work of Leonard (1978) and Silverman (1982).

Penalized likelihood method estimates function of interest, say g, by the minimizer of a score of the form

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

where $L(g|\text{data})$, usually a minus log likelihood, measures the goodness-of-fit of g to the data, $J(g)$ (≥ 0) measures the roughness of g, and the so-called smoothing parameter λ (≥ 0) controls the tradeoff. The minimizer of (1.1) is effectively the maximum likelihood estimate subject to a (soft) constraint $J(g) \leq \rho$ for some $\rho \leq 0$. For the constraint to be effective for noise reduction, the null space of $J(g)$ should have a finite dimension.

Two intrinsic constraints a probability density has to satisfy is that it is nonnegative (positivity) and that it integrates to one (unity). Assuming $f(x,y) > 0$ on the domain, the logistic density transform $f = e^g / \int e^g$ (cf. Leonard 1978) takes care of both constraints, but the many-to-one feature of the transform in usual function spaces is often inconvenient for theoretical analysis and numerical computation. For the estimation of the joint density $f(x,y)$, Gu and Qiu (1993) propose a simple surgery on usual function spaces to make the transform one-to-one. For the estimation of the conditional density $f(y|x)$, further surgery is needed. The idea can most conveniently be explained in the context of analysis of variance (ANOVA) decomposition of multivariate functions.

An ANOVA decomposition for a bivariate function is expressed as $g(x,y) = g_\emptyset + g_x(x) + g_y(y) + g_{x,y}(x,y)$, where $g_\emptyset$ is a constant, g_x and g_y are functions of one variable called the main effects, and $g_{x,y}$ is called the interaction. For the decomposition to be uniquely defined, certain side conditions have to be enforced on g_x, g_y, and $g_{x,y}$; for example, one may set $\int_{\mathcal{X}} g_x = \int_{\mathcal{Y}} g_y = \int_{\mathcal{X}} g_{x,y} = \int_{\mathcal{Y}} g_{x,y} = 0$. A general discussion of ANOVA decomposition of multivariate functions can be found in, e.g., Gu and Wahba (1992). With a uniquely defined ANOVA decomposition of $g(x,y)$, forcing $g_\emptyset = 0$ makes $f(x,y) \leftrightarrow e^g / \int_{\mathcal{X}\times\mathcal{Y}} e^g$ one-to-one, which is the surgery suggested by Gu and Qiu (1993) for the joint density. In a similar manner, one may set $g_\emptyset + g_x = 0$ to make a logistic conditional density transform $f(y|x) \leftrightarrow e^{g(x,y)} / \int_{\mathcal{Y}} e^{g(x,y)}$ one-to-one.

With a one-to-one logistic conditional density transform, one may specialize (1.1) for the estimation of $f(y|x)$ as follows. Writing $\mathcal{H} = \{g : g(x,y) = g_y(y) + g_{x,y}(x,y)\}$ where g_y and $g_{x,y}$ satisfy side conditions required in an ANOVA decomposition, one may estimate $f(y|x)$ by $e^{g(x,y)} / \int_{\mathcal{Y}} e^{g(x,y)}$ where g minimizes

$$-\frac{1}{n}\sum_{i=1}^{n}\{g(X_i,Y_i) - \log\int_{\mathcal{Y}} e^{g(X_i,y)}\} + \frac{\lambda}{2}J(g) \tag{1.2}$$

in $\mathcal{H}$, where the divisor 2 of λ saves notation in later analysis. This procedure can be implemented via tensor product splines; details are to be found in Section 2.

A large body of literature on nonparametric conditional density estimation exists under the name of regression, of which most assume parametric models for $f(y|x)$ on the y axis. Of those do not assume any parametric form, most still operate on certain parameters of $f(y|x)$ such as conditional mean or conditional percentiles. Similar to a recent work by Stone (1991) who uses tensor product regression splines in Euclidean spaces, we use tensor product smoothing splines to estimate the whole conditional density, from which distributional parameters can be readily derived. A point worth noting is that the domains $\mathcal{X}$ and $\mathcal{Y}$ in (1.2) are generic, so the method may apply to problems on arbitrary domains. For example, with a discrete $\mathcal{Y}$ one may employ the method to conduct nonparametric multinomial regression.

The remainder of the article is organized as follows. Section 2 formally sets up the problem, conducts preliminary analysis, and presents examples. Section 3 sketches a generic asymptotic theory and Section 4 notes the computation of estimates with automatic smoothing parameters. Section 5 illustrates some applications of the method. Section 6 concludes the article with a few remarks.

2 Penalized Likelihood Estimation

We first tighten up the formulation of (1.2). For (1.2) to be well defined at $g = 0$, one has to assume a bounded $\mathcal{Y}$, which presumably covers the observed Y_i; with unbounded or unknown natural support for Y the estimation shall be interpreted as that of conditional distribution of $Y|(Y \in \mathcal{Y})$. Penalty functional $J(g)$ in penalized likelihood estimation is usually taken as a quadratic form with a null space of dimension smaller than the sample size n. For (1.2) to be sensible for the estimation of $f(y|x)$, $J(g)$ should annihilate functions of x alone because $g(x, y)$ and $g(x, y) + h(x)$ for any $h(x)$ leads to identical $f(y|x)$ by the logistic conditional density transform. One may try to minimize (1.2) over all functions satisfying $J(g) < \infty$, but functions differ by a function of variable x alone are equivalent to each other, and for theoretical and computational convenience we shall allow one and only one member of each equivalent class in the estimation. This can be done by forcing $A_y g(x, y) = 0$, where A_y is an "averaging operator" acting on variable y which preserves functions of x alone; among examples of A_y are $A_y g = \int_{\mathcal{Y}} g / \int_{\mathcal{Y}} 1$ and $A_y g = g(x, y_0)$, $y_0 \in \mathcal{Y}$. Now let $\mathcal{H} = \{g : A_y g = 0, J(g) < \infty\}$ and $J_\perp = \{g : A_y g = 0, J(g) = 0\}$. $J(g)$ forms a natural square (semi) norm in $\mathcal{H}$, and supplemented by a norm in $J_\perp$, makes $\mathcal{H}$ a Hilbert space. Evaluation appears in the likelihood part of (1.2), and it shall be assumed that evaluation is continuous in $\mathcal{H}$.

Write $t = (x, y)$ and $\mathcal{T} = \mathcal{X} \times \mathcal{Y}$. A Hilbert space in which evaluation is continuous is called a reproducing kernel Hilbert space (RKHS) possessing a reproducing kernel (RK) $R(\cdot, \cdot)$, a nonnegative definite bivariate function on $\mathcal{T}$, such that $R(t, \cdot) = R(\cdot, t) \in \mathcal{H}$, $\forall t \in \mathcal{T}$, and $\langle R(t, \cdot), g(\cdot) \rangle = g(t)$ (the reproducing property), $\forall g \in \mathcal{H}$, where $\langle \cdot, \cdot \rangle$ is the inner product in $\mathcal{H}$. As a matter of fact, starting from any nonnegative definite function $R(\cdot, \cdot)$ on the domain, one can construct a RKHS $\mathcal{H} = \text{span}\{R(t, \cdot), \forall t \in \mathcal{T}\}$ with an inner product satisfying $\langle R(t, \cdot), R(s, \cdot) \rangle = R(t, s)$, which has $R(\cdot, \cdot)$ as its RK. The inner product (hence the norm) and the RK determine each other uniquely. Details can be found in Aronszajn (1950); see also Wahba (1990, Chapter 1). With J as a square seminorm, $\mathcal{H}$ can be decomposed as $\mathcal{H} = \mathcal{H}_J \oplus J_\perp$, where $\mathcal{H}_J = \{g : g \in \mathcal{H}, J(g) \in (0, \infty)\}$ is a RKHS with a square norm J and an associated RK R_J. Note that the null space norm does not appear in (1.2). The estimate is determined by the data (X_i, Y_i) and the model implied by a basis of $J_\perp$, the RK R_J, and the smoothing parameter λ. $J(g)$ is intuitive for the perception of smoothness implied by the model, whereas R_J and a basis of $J_\perp$ are the only things needed for computation. On a product domain such as we have here, it is often easier to first construct R_J then to possibly derive explicit form of the associated $J(g)$, than to do things the other way; examples follow shortly.

Theorem 2.1 *Assume that the RK of $\mathcal{H}$ is bounded on $\mathcal{Y}$ for any fixed $x \in \mathcal{X}$. If the minimizer $\hat{g}$ of (1.2) exists in $J_\perp$, then it uniquely exists in $\mathcal{H}$.*

Proof: By Theorem 4.1 of Gu and Qiu (1993), it suffices to show that $\log \int_{\mathcal{Y}} e^{g(x,y)}$ is continuous and strictly convex in $\mathcal{H}$ for any given x. Continuity follows the continuity of evaluation, and boundedness of RK and Riemann sum approximation of $\int_{\mathcal{Y}}$ if necessary. Convexity follows Hölder's inequality, note that $\log \int_{\mathcal{Y}} e^{\alpha g + \beta h} \leq \alpha \log \int_{\mathcal{Y}} e^g + \beta \log \int_{\mathcal{Y}} e^h$ for $\alpha, \beta > 0$, $\alpha + \beta = 1$, where the equality holds only when $e^g \propto e^h$ on $\{x\} \times \mathcal{Y}$, which amounts to $g = h$ in $\mathcal{H}$ with $A_y g = 0$. □

We now derive loss functions for the assessment of estimation precision. Conditional on x, the symmetrized Kullback-Leibler between $e^g / \int_{\mathcal{Y}} e^g$ and $e^h / \int_{\mathcal{Y}} e^h$ is $\text{SKL}(g, h|x) = \mu_g(g - h|x) -$

$\mu_h(g-h|x)$ where $\mu_g(h|x) = \int_{\mathcal{Y}} he^g / \int_{\mathcal{Y}} e^g$. Observing $Y|X$ from $e^{g_0}/\int_{\mathcal{Y}} e^{g_0}$ and X from $f(x)$, $\text{SKL}(g,g_0) = \int_{\mathcal{X}} \text{SKL}(g,g_0|x)f(x)$ appears appropriate for assessing the performance of g as an estimate of g_0. A first order Taylor expansion of $\mu_{g_0+\alpha f}(h|x)$ in α at $\alpha = 0$ gives $\mu_{g_0}(h|x)+\alpha v(h,f|x)$ where $v(h,f|x) = v_{g_0}(h,f|x) = \mu_{g_0}(hf|x) - \mu_{g_0}(h|x)\mu_{g_0}(f|x)$, and by plugging in $h = f = g - g_0$ and $\alpha = 1$ one obtains a quadratic distance $V(g-g_0) = \int_{\mathcal{X}} v(g-g_0|x)f(x)$ which approximates $\text{SKL}(g-g_0)$ for g near g_0, where $v(h|x) = v(h,h|x)$. Asymptotic convergence rates of $\hat{g}$ in $\text{SKL}(\hat{g},g_0)$ and $V(\hat{g}-g_0)$ will be established in Section 3 under certain conditions.

The rest of the section are examples.

Example 2.1 *Tensor product linear splines on* $[0,1]^2$. We start with a construction of RKHS on $[0,1]$. A possible roughness functional for one dimensional smoothing on $[0,1]$ is $\int_0^1 \dot{g}^2$, which is a square semi norm in $\{g : \int_0^1 \dot{g}^2 < \infty\}$ with a null space $\{1\}$. Imposing a side condition, say $\int_0^1 g = 0$ or $g(0) = 0$, $\int_0^1 \dot{g}^2$ can be made a square norm in the reduced space and an RK can be derived. Two commonly used configurations follow. In $\{g : \int_0^1 \dot{g}^2 < \infty, \int_0^1 g = 0\}$, the RK associated with square norm $\int_0^1 \dot{g}^2$ is $R_{l1}(x_1,x_2) = k_1(x_1)k_1(x_2) + k_2(|x_1-x_2|)$, where $k_\nu = B_\nu/\nu!$ and B_ν is the νth Bernoulli polynomial (cf. Craven and Wahba 1979). In $\{g : \int_0^1 \dot{g}^2 < \infty, g(0) = 0\}$ the RK is $R_{l2}(x_1,x_2) = \min(x_1,x_2)$. It can be verified that $\int_0^1 R_{l1}(x_1,x_2)dx_2 = 0$ and $R_{l2}(x_1,0) = 0$. $R_0(x_1,x_2) = 1$ is an RK for $\{1\}$. $R_0 + R_{l1}$ and $R_0 + R_{l2}$ generate RKHS with square norms $(\int_0^1 g)^2 + \int_0^1 \dot{g}^2$ and $g^2(0) + \int_0^1 \dot{g}^2$, respectively, and they represent one-way ANOVA with different side conditions. Estimates with $J(g) = \int_0^1 \dot{g}^2$ are linear splines.

With nonnegative definite functions R^x and R^y on $\mathcal{X}$ and $\mathcal{Y}$, respectively, $R((x_1,y_1),(x_2,y_2)) = R^x(x_1,x_2)R^y(y_1,y_2)$ is nonnegative definite on $\mathcal{X} \times \mathcal{Y}$ (cf. Aronszajn 1950). This fact serves as a convenient device for the construction of RKHSs on product domains. From the marginals $R_0 + R_{l1}$ or $R_0 + R_{l2}$, one readily obtains RKHSs on $[0,1]^2$ with ANOVA decompositions built-in. For example, $R^x_0 R^y_0 + R^x_{l1} R^y_0 + R^x_0 R^y_{l1} + R^x_{l1} R^y_{l1}$ generates $g_\emptyset + g_x + g_y + g_{x,y}$ with side conditions $\int_0^1 g_x = \int_0^1 g_y = \int_0^1 g_{x,y}dx = \int_0^1 g_{x,y}dy = 0$, and replacing R_{l1} by R_{l2} generates the same expression but with side conditions $g_x(0) = g_y(0) = g_{x,y}(0,y) = g_{x,y}(x,0) = 0$. Cutting off $g_\emptyset$ and g_x, one obtains an $\mathcal{H}$ for the purpose of (1.2).

Specifically, an RK $R_J = \theta_1 R^y_{l1} + \theta_2 R^x_{l1} R^y_{l1}$ generates an RKHS $\mathcal{H} = \{g : \int_0^1 g dy = 0, J(g) < \infty\}$ where $J(g) = \theta_1^{-1} \int_0^1 (\int_0^1 (\partial g/\partial y)dx)^2 dy + \theta_2^{-1} \int_0^1 \int_0^1 (\partial^2 g/\partial x \partial y)^2 dxdy$, and replacing R^y_{l1} by R^y_{l2} only changes the side condition in $\mathcal{H}$ but not $J(g)$. Similarly, $R_J = \theta_1 R^y_{l2} + \theta_2 R^x_{l2} R^y_{l2}$ generates an RKHS

$\mathcal{H} = \{g : g(x,0) = 0, J(g) < \infty\}$ where $J(g) = \theta_1^{-1} \int_0^1 (\partial g/\partial y)^2(0,y)dy + \theta_2^{-1} \int_0^1 \int_0^1 (\partial^2 g/\partial x \partial y)^2 dxdy$, and replacing R_{l2}^y by R_{l1}^y only changes the side condition in $\mathcal{H}$ but not $J(g)$. Extra smoothing parameters θ_β (> 0) to be selected by the data are attached to terms of R_J because the scalings of individual terms are not comparable. There are no clearly separable finite dimensional parts in g_y and $g_{x,y}$ and one may set $J_\perp = \{0\}$. Note that the two $J(g)$ imply slightly different notions of smoothness due to the different side conditions on the x axis which affects the break-up of $g_y + g_{x,y}$. The derivations of $J(g)$ are straightforward but tedious, which we omit. The minimizer $\hat{g}$ of (1.2) always uniquely exists in this setup. □

Example 2.2 *Tensor product cubic splines on* $[0,1]^2$. We again start with a construction on $[0,1]$. The most commonly used roughness functional for one dimensional smoothing is $\int_0^1 \ddot{g}^2$, which is a square semi norm in $\{g : \int_0^1 \ddot{g}^2 < \infty\}$ with a null space of linear polynomials $\{1, (\cdot)\}$. Imposing a pair of side conditions, say $\int_0^1 g = \int_0^1 \dot{g} = 0$ or $g(0) = \dot{g}(0) = 0$, $\int_0^1 \ddot{g}^2$ can be made a square norm in the reduced space and an RK can be derived. Two commonly used configurations follow. In $\{g : \int_0^1 \ddot{g}^2 < \infty, \int_0^1 g = \int_0^1 \dot{g} = 0\}$, the RK associated with square norm $\int_0^1 \ddot{g}^2$ is $R_{c1}(x_1, x_2) = k_2(x_1)k_2(x_2) - k_4(|x_1 - x_2|)$ with $k_\nu = B_\nu/\nu!$ scaled Bernoulli polynomials (cf. Craven and Wahba 1979); accompanying RKs $R_0 = 1$ and $R_{\pi 1}(x_1, x_2) = (x_1 - .5)(x_2 - .5)$ generate $\{1\}$ and $\{(\cdot - .5)\}$ with square norms $(\int_0^1 g)^2$ and $(\int_0^1 \dot{g})^2$, respectively, and the tensor sum of the three subspaces forms an RKHS. In $\{g : \int_0^1 \ddot{g}^2 < \infty, g(0) = \dot{g}(0) = 0\}$ the RK is $R_{c2}(x_1, x_2) = \int_0^1 (x_1 - u)_+ (x_2 - u)_+ du$ where $(\cdot)_+ = \max(0, \cdot)$; accompanying RKs $R_0 = 1$ and $R_{\pi 1}(x_1, x_2) = x_1 x_2$ generate $\{1\}$ and $\{(\cdot)\}$ with square norms $g^2(0)$ and $\dot{g}^2(0)$, respectively, and the tensor sum of the three subspaces forms another RKHS with a different norm. $R_0 + (R_{\pi 1} + R_{c1})$ and $R_0 + (R_{\pi 2} + R_{c2})$ represent one-way ANOVA with different side conditions. Estimates with $J(g) = \int_0^1 \ddot{g}^2$ are cubic splines.

Using marginals $R_0 + (R_{\pi 1} + R_{c1})$ or $R_0 + (R_{\pi 2} + R_{c2})$, one can paste up RKHSs on $[0,1]^2$ with up to nine tensor sum subspaces. For example, $\theta_{0,0} R_0^x R_0^y + (\theta_{\pi,0} R_{\pi 1}^x R_0^y + \theta_{c,0} R_{c1}^x R_0^y) + (\theta_{0,\pi} R_0^x R_{\pi 1}^y + \theta_{0,c} R_0^x R_{c1}^y) + (\theta_{\pi,\pi} R_{\pi 1}^x R_{\pi 1}^y + \theta_{\pi,c} R_{\pi 1}^x R_{c1}^y + \theta_{c,\pi} R_{c1}^x R_{\pi 1}^y + \theta_{c,c} R_{c1}^x R_{c1}^y)$, $\theta_\beta \geq 0$, generates $g_\emptyset + g_x + g_y + g_{x,y}$ with side conditions $\int_0^1 g_x = \int_0^1 g_y = \int_0^1 g_{x,y} dx = \int_0^1 g_{x,y} dy = 0$, and replacing $R_{\pi 1} + R_{c1}$ by $R_{\pi 2} + R_{c2}$ yields different side conditions $g_x(0) = g_y(0) = g_{x,y}(0,y) = g_{x,y}(x,0) = 0$. A $\theta_\beta = 0$ eliminates the corresponding subspace from the space. The square norm in a pasted RKHS with an RK $\sum_\beta \theta_\beta R_\beta$ is $\sum_\beta \theta_\beta^{-1} J_\beta$, where J_β is the square norm in the space generated by R_β.

For the purpose of (1.2), one should set $\theta_{0,0} = \theta_{\pi,0} = \theta_{c,0} = 0$ and use a penalty of the form

$J(g) = \sum_{\beta \in \{0,\pi,c\} \times \{\pi,c\}} \theta_\beta^{-1} J_\beta$. Following common practice, one may put the polynomials into $J_\perp$ by setting $\theta_{0,\pi} = \theta_{\pi,\pi} = \infty$ in $J(g)$, and in turn $R_{0,\pi}$ and $R_{\pi,\pi}$ will not appear in the expression of the RK R_J in $\mathcal{H} \ominus J_\perp$. Different configurations on the x axis still imply different notions of smoothness. Furthermore, different configurations on the y margin, which now differ not only in ANOVA side conditions but also in other aspects, also imply different notions of smoothness. We omit explicit expressions of J_β, which may be found, e.g., in Gu and Wahba (1992) under slightly different notations. Under setup with a null space $J_\perp = \{(y - .5), (x - .5)(y - .5)\}$, the minimizer $\hat{g}$ of (1.2) uniquely exists whenever the maximum likelihood estimate of the form $g(x, y) = \beta_1(y - .5) + \beta_2(x - .5)(y - .5)$ exists.

We note that marginal configurations are independent of each other. For example, one may well use cubic spline on one margin and linear spline on the other. □

Example 2.3 *Tensor product splines on $\mathcal{X} \times \{1, \cdots, K\}$.* Both domains $\mathcal{X}$ and $\mathcal{Y}$ are generic in (1.2). In particular, the response domain $\mathcal{Y}$ can be taken as a discrete set, say $\{1, \cdots, K\}$, and the method can be used to conduct regression with multinomial responses. When $K = 2$, one has yet another approach to logistic regression. For the the method to apply, one needs to construct an RKHS on the marginal domain $\{1, \cdots, K\}$ with an ANOVA decomposition built in, to cut off the constant, and to take tensor product of what is left with an RKHS on the covariate domain $\mathcal{X}$.

An function on $\{1, \cdots, K\}$ is simply a K-vector and an RK a $K \times K$ nonnegative definite matrix. The integral $\int_{\mathcal{Y}}$ may be taken as summation over the domain. Smoothing on a discrete domain is better known as shrinking, and different choices of J, or equivalently R_J, imply the shrinking of different features of vector in estimation. For example, the "length" penalty $J(g) = \boldsymbol{g}^T \boldsymbol{g}$ shrinks $\boldsymbol{g}$ towards $\mathbf{o}$, whereas the "variance" penalty $J(g) = \boldsymbol{g}^T(I - \mathbf{1}\mathbf{1}^T/K)\boldsymbol{g}$ shrinks $\boldsymbol{g}$ towards $\{\mathbf{1}\}$, where we use boldface letters for the K-vectors. The RK corresponding to the square norm $J(g) = \boldsymbol{g}^T(I - \mathbf{1}\mathbf{1}^T/K)\boldsymbol{g}$ in $\{\mathbf{1}\}^\perp$ is $R_v = (I - \mathbf{1}\mathbf{1}^T/K)$, which generates vectors satisfying the side condition $\mathbf{1}^T\boldsymbol{g} = 0$. Actually, it can be shown that the RK corresponding to a quadratic square norm $\boldsymbol{g}^T A \boldsymbol{g}$ in the column space of A is A^+, the Moore-Penrose inverse of A; see, e.g., Gu and Wahba (1992).

Following the same procedure as used in previous examples, one can easily construct tensor product RKHS on $\mathcal{X} \times \{1, \cdots, K\}$ by taking product of R_v with RK's on $\mathcal{X}$. For example, with $\mathcal{X} = [0, 1]$, one may use $R_J = \theta_\pi R^x_{\pi 1} R^y_v + \theta_c R^x_{c1} R^y_v$ and $J_\perp = \{R_v(j, \cdot)\}_{j=1}^{K-1}$ for (1.2), where for example $\theta_\pi < \infty$ implies the shrinking of the "variance" of slopes on the x axis for different y

values. If K is small, one may allow $\theta_\pi = \infty$ to put the linear term(s) into $J_\perp$.

Obviously, R_v is not the only nonnegative definite matrix which generates $\{\mathbf{1}\}^\perp$, and $\mathbf{1}^T g = 0$ is not the only choice for the side condition in an ANOVA decomposition on $\{1, \cdots, K\}$. For example, with ordinal categories one may choose to use an RK corresponding to $J(g) = \sum_{j=1}^{K-1}(g(j+1)-g(j))^2$ in $\{\mathbf{1}\}^\perp$, which shrinks the differences between adjacent categories. □

3 Asymptotic Theory

Assume $g_0 \in \mathcal{H}$ and the maximum likelihood estimate exists in $J_\perp$ so $\hat{g}$ exists. We shall establish the asymptotic convergence rates of $\hat{g}$ and those of a computable approximation. The theory runs parallel to that of Gu and Qiu (1993), and to avoid too much overlap with that article, we shall only present a sketch here. Details can easily be filled in following the lines of Gu and Qiu (1993).

Assuming $f(x) > 0$ on $\mathcal{X}$, $V(g) = \int_{\mathcal{X}} v(g|x)f(x)$ defines a square norm in $\mathcal{H} \subseteq \{g : A_y g = 0\}$ interpretable under the stochastic structure. $J(g)$ defines the notion of smoothness. A characterization of the models implied by (1.2) is via an eigenvalue analysis of J with respect to V. A bilinear form B is said to be completely continuous with respect to another bilinear form A, if for any $\epsilon > 0$, there exist finite number of linear functionals $l_1, \cdots, l_{k_\epsilon}$ such that $l_j(\eta) = 0$, $j = 1, \cdots, k_\epsilon$, implies that $B(\eta) \leq \epsilon A(\eta)$ (cf. Weinberger 1974, Section 3.3).

Assumption A.1. V is completely continuous with respect to J.

Under A.1, it can be shown that there exist $\phi_\nu \in \mathcal{H}$ and $0 \leq \rho_\nu \uparrow \infty$, $\nu = 1, 2, \cdots$, such that $V(\phi_\nu, \phi_\mu) = \delta_{\nu,\mu}$ and $J(\phi_\nu, \phi_\mu) = \rho_\nu \delta_{\nu,\mu}$, where $\delta_{\nu,\mu}$ is the Kronecker delta and $V(\cdot,\cdot)$ and $J(\cdot,\cdot)$ are the (semi) inner products associated with $V(g)$ and $J(g)$ (cf. Gu and Qiu 1993, Section 5). Since $J(g) = \sum_\nu g_\nu^2 \rho_\nu$ and $\rho_\nu \uparrow \infty$, A.1 implies that the term $\lambda J(g)$ in (1.2) for any fixed λ restricts the model space to an effectively finite dimension in terms of the V norm, which is necessary for noise reduction, and that the effective model space dimension can be expanded by letting $\lambda \to 0$ as $n \to \infty$. The rate of growth of ρ_ν quantifies the notion of smoothness implied by $J(g)$.

Assumption A.2. $\rho_\nu = c_\nu \nu^r$, where $r > 1$, $c_\nu \in (\beta_1, \beta_2)$, and $0 < \beta_1 < \beta_2 \leq \infty$.

For Examples 2.1 and 2.2, A.1 and A.2 are both satisfied, with $r = 2 - \epsilon$ and $r = 4 - \epsilon$ for linear and cubic splines, respectively, where $\epsilon > 0$ is positive but arbitrary; for Example 2.3, we note that the r in a tensor product RKHS with a mixture of discrete and continuous margins is determined by that in the partial product space of all the continuous margins. See, e.g., Utreras (1981) and Gu (1992b) for relevant technical details.

Denote by $P_\lambda(g)$ the expression in (1.2). Define $Q_\lambda(g) = -(1/n)\sum_{i=1}^n\{g(X_i,Y_i) - \mu_{g_0}(g|X_i)\} + (1/2)V(g-g_0) + (\lambda/2)J(g)$, the quadratic approximation of $P_\lambda(g)$ at g_0. Let g_1 be the minimizer of $Q_\lambda(g)$ in $\mathcal{H}$, which exists similar to $\hat{g}$. When $g_0 \in \mathcal{H}$, it can be shown that $(V+\lambda J)(g_1 - g_0) = O_p(n^{-1}\lambda^{-1/r} + \lambda)$ under A.1 and A.2 as $\lambda \to 0$ and $n\lambda^{1/r} \to \infty$ (cf. Gu and Qiu 1993, Section 5).

A few technical assumptions follow.

Assumption A.3. For g in a convex set B_0 around g_0 containing $\hat{g}$ and g_1, $\exists c_1, c_2 \in (0,\infty)$ such that $c_1 v(h|x) \le v_g(h|x) \le c_2 v(h|x), \forall x \in \mathcal{X}$.

It is easy to show that $\mu_g(g-h|x) - \mu_h(g-h|x) = v_{\alpha g + (1-\alpha)h}(g-h|x)$ for some $\alpha \in [0,1]$, so A.3 implies the equivalence of $v(g-h|x)$ and $\mathrm{SKL}(g,h|x)$, and hence of $V(g-h)$ and $\mathrm{SKL}(g,h)$, for $g,h \in B_0$.

Assumption A.4. $\exists c_3 < \infty$ such that $\int_{\mathcal{X}} v^2(\phi_\nu|x)f(x) \le c_3, \forall \nu$.

Assumption A.5. $\exists c_4 < \infty$ such that
$$\int_{\mathcal{X}}[v(\phi_\nu\phi_\mu|x) + \{\mu_{g_0}(\phi_\nu\phi_\mu|x) - \int_{\mathcal{X}}\mu_{g_0}(\phi_\nu\phi_\mu|x)f(x)\}^2]f(x) \le c_4, \forall \nu, \mu.$$

A uniform bound on the fourth moments of $\phi_\nu(X,Y)$ is sufficient for A.4 and A.5 to hold.

Theorem 3.1 *Assume $g_0 \in \mathcal{H}$. Under A.1 - A.4, as $\lambda \to 0$ and $n\lambda^{2/r} \to \infty$,* $(V+\lambda J)(\hat{g}-g_0) \sim \mathrm{SKL}(\hat{g}, g_0) = O_p(n^{-1}\lambda^{-1/r} + \lambda)$.

A sketch of the proof follows. Let $A_{g,h}(\alpha) = P_\lambda(g+\alpha h)$ and $B_{g,h}(\alpha) = Q_\lambda(g+\alpha h)$. Equating $\dot{A}_{\hat{g},\hat{g}-g_1}(0) = \dot{B}_{g_1,\hat{g}-g_1}(0) = 0$, A.3 and A.4 lead to $(c_1V + \lambda J)(\hat{g}-g_1) = (V+\lambda J)^{1/2}(\hat{g}-g_1)(V+\lambda J)^{1/2}(g_1-g_0)O_p(1)$. The theorem then follows $(V+\lambda J)(g_1-g_0) = O_p(n^{-1}\lambda^{-1/r} + \lambda)$.

The space $\mathcal{H}$ is in general infinite dimensional and $\hat{g}$ not computable. An important part of the theory is to justify a computable approximation of $\hat{g}$ in a finite dimensional space. Let $\mathcal{H}_n = J_\perp \oplus \mathrm{span}\{R_J(X_i,\cdot), i = 1,\cdots,n\}$ with R_J the RK in $\mathcal{H} \ominus J_\perp$. The minimizer $\hat{g}_n$ of (1.2) in $\mathcal{H}_n$ is an estimate to use in practice.

Theorem 3.2 *Modify A.3 to also include $\hat{g}_n$ and g_n in the convex set B_0, where g_n is the projection of $\hat{g}$ in $\mathcal{H}_n$. Under A.1 - A.5, as $\lambda \to 0$ and $n\lambda^{2/r} \to \infty$, $(V+\lambda J)(\hat{g}_n - g_0) \sim \mathrm{SKL}(\hat{g}_n, g_0) = O_p(n^{-1}\lambda^{-1/r} + \lambda)$.*

The proof of Theorem 3.2 goes as follows. First, it can be shown that $V(h) = o_p(\lambda J(h))$, $\forall h \in \mathcal{H} \ominus \mathcal{H}_n$. This, $\dot{A}_{\hat{g},\hat{g}-g_n}(0) = 0$, and Theorem 3.1 then yield $V(\hat{g} - g_n) = o_p(n^{-1}\lambda^{-1/r} + \lambda)$ and $\lambda J(\hat{g} - g_n) = O_p(n^{-1}\lambda^{-1/r} + \lambda)$. Equating $\dot{A}_{\hat{g}_n,\hat{g}_n-g_n}(0) = \dot{A}_{\hat{g},\hat{g}_n-\hat{g}}(0) = 0$, it can further be shown that $(V+\lambda J)(\hat{g}_n - g_n) = O_p(n^{-1}\lambda^{-1/r} + \lambda)$. The theorem then follows.

4 Computation

Write $\xi_i = R_J((X_i, Y_i), \cdot)$ and $J_\perp = \{\phi_\nu\}_{\nu=1}^M$. A function in $\mathcal{H}_n$ has an expression $g = \sum_{i=1}^n c_i \xi_i + \sum_{\nu=1}^M d_\nu \phi_\nu = \boldsymbol{\xi}^T \mathbf{c} + \boldsymbol{\phi}^T \mathbf{d}$, where $\boldsymbol{\xi}$ and $\boldsymbol{\phi}$ are vectors of functions and $\mathbf{c}$ and $\mathbf{d}$ are vectors of coefficients. Fixing smoothing parameters, $\hat{g}_n$ can be calculated via minimizing

$$-\frac{1}{n}\mathbf{1}^T(Q\mathbf{c} + S\mathbf{d}) + \frac{1}{n}\sum_{i=1}^n \log \int_{\mathcal{Y}} \exp\{\boldsymbol{\xi}_i^T \mathbf{c} + \boldsymbol{\phi}_i^T \mathbf{d}\} + \frac{\lambda}{2}\mathbf{c}^T Q \mathbf{c} \tag{4.1}$$

with respect to $\mathbf{c}$ and $\mathbf{d}$, where Q is $n \times n$ with (i,j)th entry $\xi_i(X_j, Y_j) = R_J((X_i, X_i), (X_j, Y_j))$, S is $n \times M$ with (i,ν)th entry $\phi_\nu(X_i, Y_i)$, $\boldsymbol{\xi}_i$ is $n \times 1$ with jth entry $\xi_j(X_i, y)$, and $\boldsymbol{\phi}_i$ is $M \times 1$ with νth entry $\phi_\nu(X_i, y)$. Substituting the empirical distribution for $f(x)$, we write $\mu_g(h) = (1/n)\sum_{i=1}^n \mu_g(h|X_i)$ and $V_g(h, f) = (1/n)\sum_{i=1}^n v_g(h, f|X_i)$. From an estimate $\tilde{g} = \boldsymbol{\xi}^T \tilde{\mathbf{c}} + \boldsymbol{\phi}^T \tilde{\mathbf{d}}$, the one-step Newton update for minimizing (4.1) can be shown to satisfy

$$\begin{pmatrix} V_{\xi,\xi} + \lambda Q & V_{\xi,\phi} \\ V_{\phi,\xi} & V_{\phi,\phi} \end{pmatrix} \begin{pmatrix} \mathbf{c} \\ \mathbf{d} \end{pmatrix} = \begin{pmatrix} Q\mathbf{1}/n - \boldsymbol{\mu}_\xi + V_{\xi,g} \\ S^T \mathbf{1}/n - \boldsymbol{\mu}_\phi + V_{\phi,g} \end{pmatrix}, \tag{4.2}$$

where $\boldsymbol{\mu}_\xi = \mu_{\tilde{g}}(\boldsymbol{\xi})$, $\boldsymbol{\mu}_\phi = \mu_{\tilde{g}}(\boldsymbol{\phi})$, $V_{\xi,\xi} = V_{\tilde{g}}(\boldsymbol{\xi}, \boldsymbol{\xi}^T)$, $V_{\xi,\phi} = V_{\tilde{g}}(\boldsymbol{\xi}, \boldsymbol{\phi}^T)$, $V_{\phi,\phi} = V_{\tilde{g}}(\boldsymbol{\phi}, \boldsymbol{\phi}^T)$, $V_{\xi,g} = V_{\tilde{g}}(\boldsymbol{\xi}, \tilde{g})$, and $V_{\phi,g} = V_{\tilde{g}}(\boldsymbol{\phi}, \tilde{g})$.

With varying smoothing parameters, (4.2) 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 with automatic multiple smoothing parameters has been implemented in Gu (1993b), where its applications to joint density estimation and covariate dependent hazard estimation are illustrated. The algorithm is directly applicable to conditional density estimation, with performance measured by estimated proxies of $V(g - g_0)$ or $\mathrm{SKL}(g, g_0)$. Motivation, intuition, and technical details can be found in Gu (1993a, b).

5 Applications

In this section, we illustrate some applications of the technique in data analysis.

5.1 Penny thickness data: Known discontinuity

The data are thickness in mils of a sample of 90 U.S. Lincoln pennies listed in Scott (1992, Appendix B.4). Two pennies from each year between 1945 to 1989 were measured. I mapped $\mathcal{X} \times \mathcal{Y} = [1944.5, 1989.5] \times [49, 61]$ onto $[0,1]^2$, and used the tensor product cubic spline of Example 2.2 with side conditions $\int_{\mathcal{X}} g_x = \int_{\mathcal{Y}} g_y = \int_{\mathcal{X}} g_{x,y} = \int_{\mathcal{Y}} g_{x,y}$, with $R_J = \theta_{0,c}R^y_{c1} + \theta_{\pi,c}R^x_{\pi 1}R^y_{c1} + \theta_{c,\pi}R^x_{c1}R^y_{\pi 1} + \theta_{c,c}R^x_{c1}R^y_{c1}$ and $J_\perp = \{(y - .5), (x - .5)(y - .5)\}$. The performance-oriented iteration effectively shrank the terms $\theta_{\pi,c}R^x_{\pi 1}R^y_{c1}$ and $\theta_{c,c}R^x_{c1}R^y_{c1}$. The automatic estimate of $f(y|x)$ is sketched in the left frame of Figure 5.1, where the solid line marks the conditional medians, the dashed lines the conditional quartiles, and the horizontal dotted lines the conditional 5th and 95th percentiles. The data are superimposed as circles, with the x coordinate slightly perturbed to unmask a few overlaps. The estimate is under the assumption of smoothness of log conditional density on both axes, despite the apparent abrupt downward shift of thickness from 1974 to 1975. A vertical dotted line is superimposed to mark the break.

A usual approach to regression with known breaks is to add jumps at breaks, known as the partial spline technique (cf. Wahba 1990). We shall try an adaptation here for conditional density estimation. To keep symmetry between the two sides of the break, the marginal RKHS on the x axis is to be generated by $R_{0l} + R_{0u} + R_{\pi 1} + R_{c1}$, where $R_{0l} = I_{[x_1 \in L]}I_{[x_2 \in L]}$ and $R_{0u} = I_{[x_1 \in U]}I_{[x_2 \in U]}$ generate window functions $\{I_{[x \in L]}\}$ and $\{I_{[x \in U]}\}$, respectively, with $L = [0, 2/3]$ and $U = (2/3, 1]$. An ANOVA decomposition is no longer available in this construction, but we do not really need one on the x axis. Taking tensor product with $R_{\pi 1} + R_{c1}$ on the y axis, we have a configuration with $R_J = \theta_{l,c}R^x_{0l}R^y_{c1} + \theta_{u,c}R^x_{0u}R^y_{c1} + \theta_{\pi,c}R^x_{\pi 1}R^y_{c1} + \theta_{c,\pi}R^x_{c1}R^y_{\pi 1} + \theta_{c,c}R^x_{c1}R^y_{c1}$ and $J_\perp = \{I_{[x \in L]}(y - .5), I_{[x \in U]}(y - .5), (x - .5)(y - .5)\}$. The performance-oriented iteration effectively shrank the terms $\theta_{\pi,c}R^x_{\pi 1}R^y_{c1}$, $\theta_{c,\pi}R^x_{c1}R^y_{\pi 1}$, and $\theta_{c,c}R^x_{c1}R^y_{c1}$. The automatic estimate with a break built-in this way is

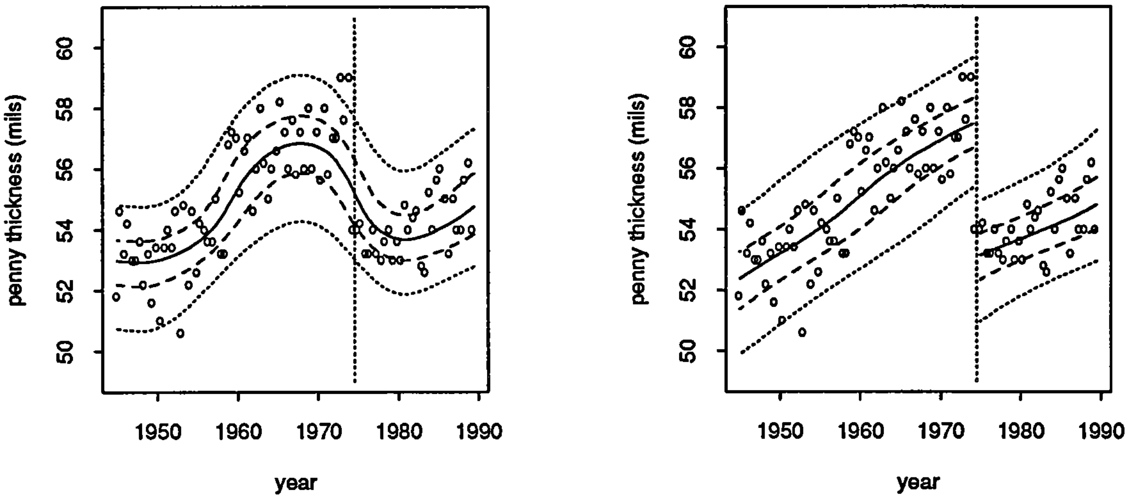

Figure 5.1: Penny Thickness Data. Left: continuous model; right: model with break. Solid lines are conditional medians, dashed lines quartiles, and dotted lines 5th and 95th percentiles. Circles are data and vertical dotted line position of the break.

sketched in the right frame of Figure 5.1 in a manner similar to the left frame.

The two configurations took about 80 and 90 cpu minutes to compute, respectively, on an IBM-RS6000.

5.2 Heart disease data: Logistic and multinomial regression

The data were collected by Dr. Robert Detrano at Cleveland Clinic Foundation on 303 patients, and taken from the UCI Repository of Machine Learning Databases (cf. Murphy and Aha 1992). There are 76 entries in the covariate list, of which only 13 were ever used by machine learning researchers. The response is diagnosis of heart disease. After preliminary analysis, I chose to model diagnosis on 3 (derived) variables: chest pain type (X_1), maximum heart rate achieved (X_2), and ST depression induced by exercise relative to rest (X_3). X_1 has four categories of typical angina, atypical angina, non-anginal pain, and asymptomatic; I lumped together the first three (call them symptomatic) which seem to have similar disease rates as much lower than that associated with asymptomatic chest pain. X_2 is covered by $[60, 210]$. After a transform $\log_{10}(x+1)$ to make it more evenly scattered, X_3 is covered by $[0, .86]$. There are five diagnostic categories, 0 for no disease,

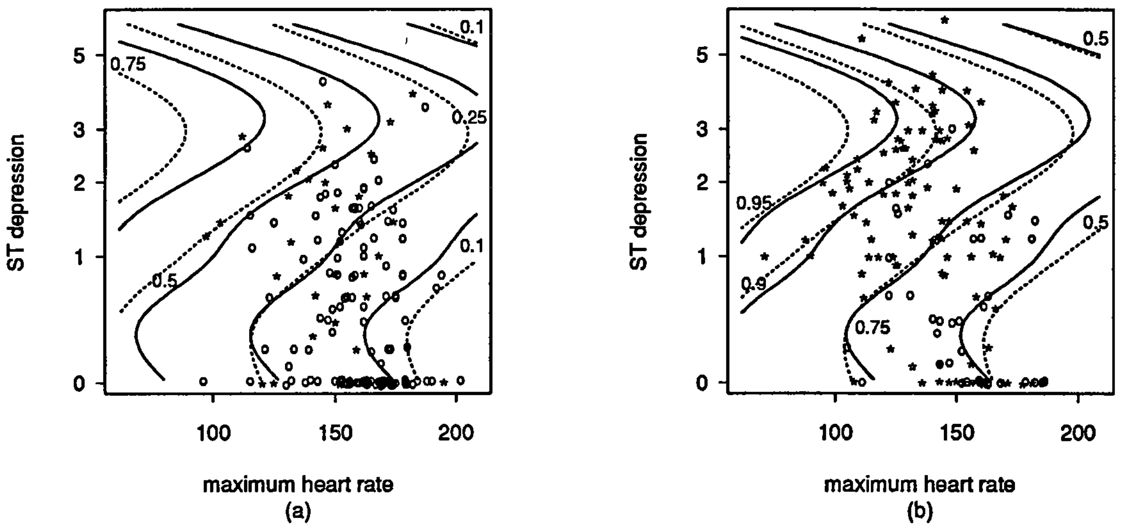

Figure 5.2: Heart Disease Data. Left: disease rates with symptomatic chest pain; right: disease rates with asymptomatic chest pain. Solid lines are estimates from logistic model and dotted lines estimates from multinomial model. Circles are healthy patients and stars disease patients.

and 1 through 4 for angiographic disease status. I shall present two parallel analyses, one with disease status aggregated and one with them separate as in the original data. The former is logistic regression and the latter is multinomial regression.

The x axis now has three dimensions, one binary and two continuous. After mapping $[60, 210] \times [0, .86]$ onto $[0, 1]^2$, one may use the tensor product cubic spline of Example 2.2 on the product domain of the two continuous covariates. To incorporate the binary covariate one may take the tensor product of $\mathbf{1}\mathbf{1}^T + R_v$ with RK's on the (product) continuous domain, but since a simple constant shift may suffice in the context, I chose to cut off all but one product term which involves $R_v^{x_1}$, that of R_v with constant, $R_v^{x_1} R_0^{x_2} R_0^{x_3}$, or effectively $R_v^{x_1}$ itself; this is the same as using the partial spline technique to add a term. The RKs on the covariate domain is thus taken as $1 + R_v^{x_1} + R_{\pi 1}^{x_2} + R_{c1}^{x_2} + R_{\pi 1}^{x_3} + R_{c1}^{x_3} + R_{\pi 1}^{x_2} R_{\pi 1}^{x_3} + R_{c1}^{x_2} R_{\pi 1}^{x_3} + R_{\pi 1}^{x_2} R_{c1}^{x_3} + R_{c1}^{x_2} R_{c1}^{x_3}$. On the y axis one may simply use R_v as the RK. Taking tensor product of RKHSs on $\mathcal{X}$ and $\mathcal{Y}$, we shall use $R_J = \theta_{\pi,0} R_{\pi 1}^{x_2} R_v^y + \theta_{c,0} R_{c1}^{x_2} R_v^y + \theta_{0,\pi} R_{\pi 1}^{x_3} R_v^y + \theta_{0,c} R_{c1}^{x_3} R_v^y + \theta_{\pi,\pi} R_{\pi 1}^{x_2} R_{\pi 1}^{x_3} R_v^y + \theta_{c,\pi} R_{c1}^{x_2} R_{\pi 1}^{x_3} R_v^y + \theta_{\pi,c} R_{\pi 1}^{x_2} R_{c1}^{x_3} R_v^y + \theta_{c,c} R_{c1}^{x_2} R_{c1}^{x_3} R_v^y$ with eight smoothing parameters, and $J_\perp = \{R_v^y(j, \cdot), R_v^{x_3}(1, \cdot) R_v^y(j, \cdot)\}_{j=1}^{K-1}$ with dimension $2(K-1)$, where K is 2 or 5, the number of diagnostic categories. Note that $R_v^{x_3}$ and R_v^y are different when

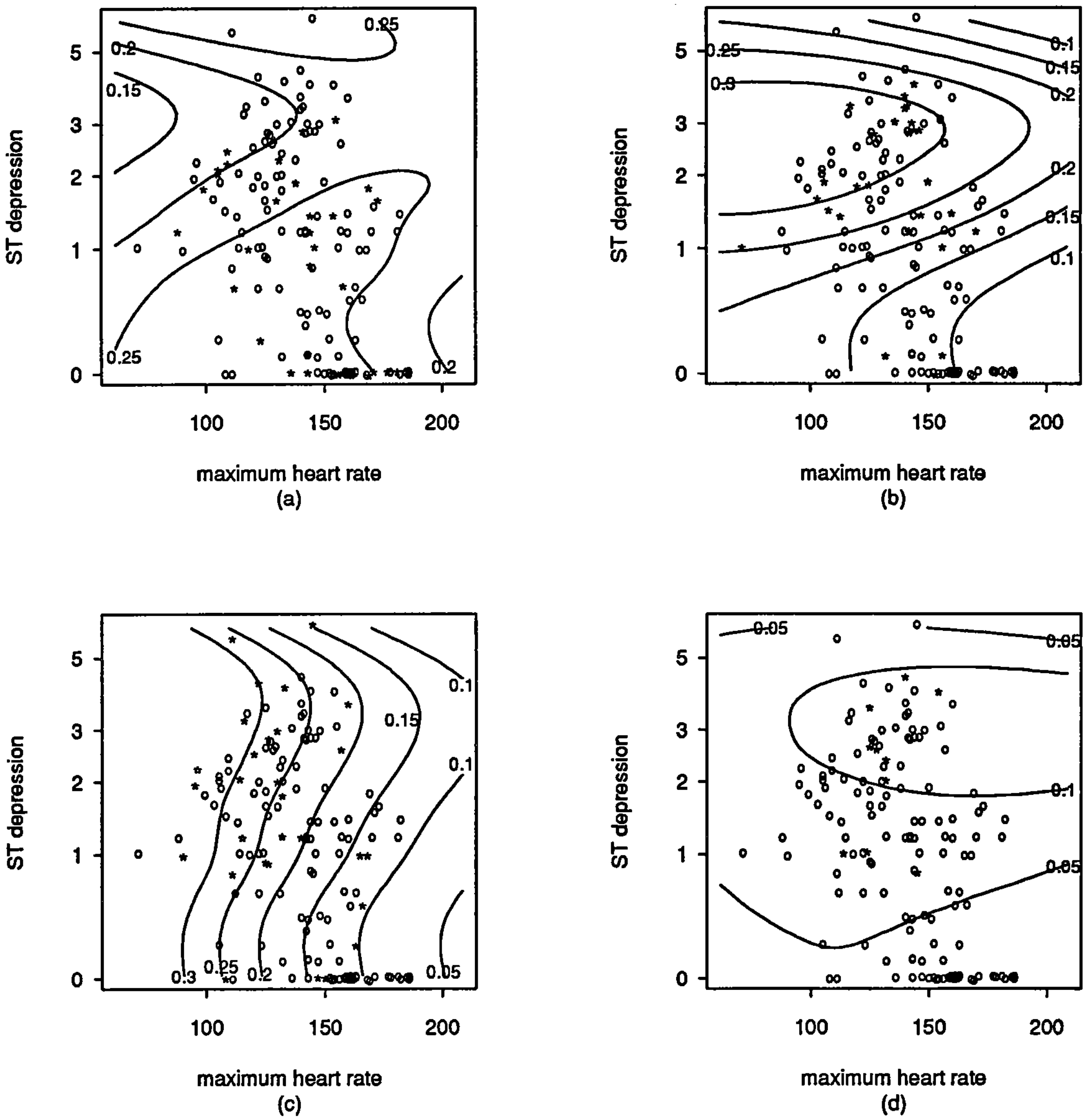

Figure 5.3: Heart Disease Data, Multinomial Model. Frames (a) – (d) contour estimated proportions of disease status 1 – 4 for patients with asymptomatic chest pain. Stars are patients in the particular disease status and circles rest of the patients.

$K = 5$ but I choose not to introduce further notation to distinguish them.

For both the logistic and the multinomial models, the performance oriented iteration effectively shrank all but two penalized terms, leaving only $\theta_{0,\pi}, \theta_{0,c} \neq 0$. The estimated disease rates are contoured in Figure 5.2, where the solid lines are from the logistic model and dotted lines aggregated from the multinomial model. Data are superimposed as circles (no disease) or stars (disease). It can be seen that the estimates from the two models agree well in data-dense area. Estimated individual proportions of the four disease categories from the multinomial model are similarly contoured for patients with asymptomatic chest pain in the four frames of Figure 5.3, where for each category the superimposed stars are patients in the corresponding status and the circles rest of the patients.

The logistic and multinomial estimates took about 112 and 165 cpu minutes to compute, respectively, on an IBM-RS6000.

6 Discussion

Research on graphical models, or density estimation with various conditional independence structures, has been rather active in recent literature, with much of the recent results focusing on the derivation of parametric distribution families for mixtures of continuous and discrete random variables; see, e.g., Wermuth and Lauritzen (1990) and Whittaker (1990) and references therein. With generic domains $\mathcal{X}$ and $\mathcal{Y}$ in (1.2), the technique presented in this article seems to pose a viable approach to nonparametric estimation of graphical models, particularly the so-called graphical chain models, a sequence of conditional distributions, possibly with a mixture of continuous and discrete variables. It is relatively straightforward to fit models with known independence structures, provided an automatic smoothing parameter selection is successful. It appears much more difficult, however, to infer independence structures from the data in a nonparametric analysis. The computational availability of nonparametric graphical models fulfills the prerequisite for research in this direction.

When $K = 2$ in Example 2.3, it is easy to verify that the estimation via (1.2) using R_v on the y axis is equivalent to that via the standard penalized likelihood logistic regression procedure as studied by O'Sullivan, Yandell, and Raynor (1986), among others, of course with the same RKHS configuration on the x axis. The smoothing parameter selection as implemented in the algorithms of Gu (1993a, b), however, is similar to but technically different from that of Gu (1992a) designed for a computation scheme in which the penalized likelihood problem is solved via a sequence of penalized weighted least squares problems. Simulation study is yet to be conducted to compare the two methods for smoothing parameter selection in logistic regression.

Acknowledgements

This research was supported by a Summer Faculty Grant from Purdue Research Foundation and a Grant from National Science Foundation.

References

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

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.

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

Gu, C. (1992a). Cross validating non Gaussian data. *J. Comput. Graph. Statist.* **1**, 169 – 179.

—— (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. *J. Amer. Statist. Assoc.* **88**, 495 – 504.

—— (1993b). Structural multivariate function estimation: Some automatic density and hazard estimates. Technical Report 93-28, Purdue University, Dept. of Statistics.

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

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

Murphy, P. M. and Aha, D. W. (1992). *UCI Repository of machine learning databases* (Machine-readable data repository). University of California–Irvine, Dept. of Information and Computer Science. Available by anonymous ftp from `ics.uci.edu` in directory `pub/machine-learning-databases`.

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.

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

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. (1991). Multivariate log-spline conditional models. Technical Report 320, University of California–Berkeley, Dept. of Statistics.

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

Weinberger, H. F. (1974). *Variational Methods for Eigenvalue Approximation.* CBMS–NSF Regional Conference Series in Applied Mathematics, Vol. 15. SIAM, Philadelphia.

Wermuth, N. and Lauritzen, S. L. (1990). On substantive research hypothesis, conditional independence graphes and graphical chain models. *J. Roy. Statist. Soc. Ser. B* **52**, 21 – 50.

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

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