Statistica Sinica **5**(1995), 709-726

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 on generic domains. Positivity and unity constraints for a probability density are enforced through a one-to-one logistic conditional density transform made possible by term trimming in an ANOVA decomposition of multivariate functions. The construction of models via tensor product splines is demonstrated through examples. The computation of estimates with automatic multiple smoothing parameters is also discussed. Data examples are presented to illustrate possible applications of the technique. For theoretical justification of the method, an asymptotic theory is sketched in the appendix.

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).

The penalized likelihood method estimates a 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 between goodness-of-fit and smoothness. The minimizer of (1.1) is effectively the maximum likelihood estimate subject to a (soft) constraint $J(g) \leq \rho$ for some $\rho \geq 0$. 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 are that it is nonnegative (positivity) and that it integrates to one (unity). Assuming $f(x,y) > 0$ on its 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 the 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 the 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 $\int_\mathcal{X} g_x = \int_\mathcal{Y} g_y = \int_\mathcal{X} g_{x,y} = \int_\mathcal{Y} g_{x,y} = 0$. Some general discussion of ANOVA decomposition of multivariate functions can be found in, e.g., Gu and Wahba (1993). 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 division of λ by 2 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 that do not assume any parametric form, most still operate on certain parameters of $f(y|x)$ such as the conditional mean or the 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 is applicable 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 discusses computational issues such as computable semiparametric approximations of the estimates and the automatic selection of smoothing parameters. Section 4 illustrates some applications of the method, including one with known discontinuity and one with a discrete $\mathcal{Y}$. Section 5 concludes the article with a few remarks. A generic asymptotic theory is sketched in the Appendix, with technical details similar to those in Gu and Qiu (1993) omitted.

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 the conditional distribution of $Y|(Y \in \mathcal{Y})$. The 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 that differ by a function of the 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 the 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 $A_y g = g(x, y_0)$, $y_0 \in \mathcal{Y}$. 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.

Write $t = (x, y)$ and $\mathcal{T} = \mathcal{X} \times \mathcal{Y}$. Evaluation appears in the likelihood part of (1.2), and we shall only consider $\mathcal{H}$ in which evaluation is continuous. Such a Hilbert space 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 $\mathcal{H}_J = \{g : g \in \mathcal{H}, 0 < J(g) < \infty\}$ is a RKHS with a square norm $J(g)$ and an associated RK R_J. The null space norm does not appear in (1.2), so 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. To obtain an RKHS $\mathcal{H}$ on the product domain $\mathcal{X} \times \mathcal{Y}$ which satisfies the aforementioned requirements for conditional density estimation, we shall construct a tensor product RKHS with an ANOVA decomposition built in, and then trim the function space components which represent the constant and the x main effect. Instead of specifying $J(g)$ directly, the method operates on the construction of RK's on the marginal and product domains, and the corresponding $J(g)$ only falls out after the fact if an explicit expression is at all available. Details will be spelled out in a few examples, to follow after the existence theorem below.

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 from the continuity of evaluation, and the boundedness of the RK and Riemann sum approximation of $\int_\mathcal{Y}$ if necessary. Convexity follows via Hölder's inequality, since $\log \int_\mathcal{Y} e^{\alpha g + \beta h} \le \alpha \log \int_\mathcal{Y} e^g + \beta \log \int_\mathcal{Y} e^h$ 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 $A_y g = 0$.

The rest of the section focuses on examples.

Example 2.1. *Tensor product linear splines on* $[0,1]^2$. We start with the construction of an 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 the 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's 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. $J(g) = \int_0^1 \dot{g}^2$ in one dimensional smoothing yields 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 RKHS's on product domains. From the marginals $R_0 + R_{l1}$ or $R_0 + R_{l2}$, one readily obtains RKHS's on $[0, 1]^2$ with ANOVA decompositions built-in. For example, $R_0^x R_0^y + R_{l1}^x R_0^y + R_0^x R_{l1}^y + R_{l1}^x R_{l1}^y$ 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 $R_J = \theta_1 R_{l1}^y + \theta_2 R_{l1}^x R_{l1}^y$ 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_{l1}^y by R_{l2}^y only changes the side condition in $\mathcal{H}$ but not $J(g)$. Similarly, $R_J = \theta_1 R_{l2}^y + \theta_2 R_{l2}^x R_{l2}^y$ 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 $J(g)$'s imply slightly different notions of smoothness due to the different side conditions on the x axis which affect the break-up of $g_y + g_{x,y}$. The derivations of $J(g)$ are straightforward but tedious and are therefore omitted. 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 $\{1, x\}$. 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 the 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 RK's $R_0 = 1$ and $R_{\pi 1}(x_1, x_2) = (x_1 - .5)(x_2 - .5)$ generate $\{1\}$ and $\{(x - .5)\}$ 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 RK's $R_0 = 1$ and $R_{\pi 1}(x_1, x_2) = x_1 x_2$ generate $\{1\}$ and $\{x\}$ 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. $J(g) = \int_0^1 \ddot{g}^2$ in one dimensional smoothing yields cubic splines.

Using marginals $R_0 + (R_{\pi 1} + R_{c1})$ or $R_0 + (R_{\pi 2} + R_{c2})$, one can paste up RKHS's on $[0,1]^2$ with up to nine tensor sum subspaces. For example, $\theta_{0,0} R_0^x R_0^y + (\theta_{\pi,0} R_{\pi 1}^x R_0^y + \theta_{c,0} R_{c1}^x R_0^y) + (\theta_{0,\pi} R_0^x R_{\pi 1}^y + \theta_{0,c} R_0^x R_{c1}^y) + (\theta_{\pi,\pi} R_{\pi 1}^x R_{\pi 1}^y + \theta_{\pi,c} R_{\pi 1}^x R_{c1}^y + \theta_{c,\pi} R_{c1}^x R_{\pi 1}^y + \theta_{c,c} R_{c1}^x R_{c1}^y)$ generates $g_\emptyset + g_x + g_y + g_{x,y}$ with side conditions $\int_0^1 g_x = \int_0^1 g_y = \int_0^1 g_{x,y} dx = \int_0^1 g_{x,y} dy = 0$, where $\theta_\beta \geq 0$, $\beta \in \{0, \pi, c\}^2$ are extra smoothing parameters; replacing $R_{\pi 1} + R_{c1}$ by $R_{\pi 2} + R_{c2}$ yields different side conditions $g_x(0) = g_y(0) = g_{x,y}(0,y) = g_{x,y}(x,0) = 0$. Setting a θ_β to 0 eliminates the corresponding subspace. 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 which generates $\mathcal{H}_J$. 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 the ANOVA side conditions but also in other aspects, also imply different notions of smoothness. We omit explicit expressions of J_β, some of which may be found, e.g., in Gu (1993c) under slightly different notation. Under the 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.

Note that the marginal configurations are independent of each other. For example, one may well use a cubic spline on one margin and a 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. For 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 the 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, with evaluation understood as coordinate extraction. The integral $\int_{\mathcal{Y}}$ may be taken as summation over the domain. Smoothing on a discrete domain is better known as shrinking, and the choice of the roughness functional specifies what is to be shrunk in estimation. For example, the "variance" penalty $J(\boldsymbol{g}) = \boldsymbol{g}^T(I - \mathbf{1}\mathbf{1}^T/K)\boldsymbol{g}$ shrinks $\boldsymbol{g}$ towards $\{\mathbf{1}\}$, where we use boldface letters for the K-vectors. The RK corresponding to the square norm $\boldsymbol{g}^T(I - \mathbf{1}\mathbf{1}^T/K)\boldsymbol{g}$ in $\{\mathbf{1}\}^{\perp}$ is $R_v = (I - \mathbf{1}\mathbf{1}^T/K)$, whose columns 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.

Following the same procedure as used in previous examples, one can easily construct a tensor product RKHS on $\mathcal{X} \times \{1, \cdots, K\}$ by taking the product of R_v with RK's on $\mathcal{X}$. For example, with $\mathcal{X} = [0, 1]$, one may use for (2.1) $R_J = \theta_\pi R^x_{\pi 1} R^y_v + \theta_c R^x_{c1} R^y_v$, where for clarity we note that $R^y_v(y_1, y_2)$ is the (y_1, y_2)th entry of the matrix $(I - \mathbf{1}\mathbf{1}^T/K)$, and $J_\perp = \{R_v(y, j)\}_{j=1}^{K-1} = \{b_j(y)\}_{j=1}^{K-1}$, where $R_v(y, j) = b_j(y)$ is the jth column of $(I - \mathbf{1}\mathbf{1}^T/K)$. Such an R_J generates an RKHS $\mathcal{H}_J = \{g(x, y) : g(x, y) = \beta_y(x - .5) + h(x, y), \sum_{y=1}^K \beta_y = 0, \sum_{y=1}^K h(x, y) = 0, \int_0^1 \ddot{h}^2_{xx}(x, y)dx < \infty, \int_0^1 h(x, y)dx = \int_0^1 \dot{h}_x(x, y)dx = 0\}$ with a square norm $J(g) = \theta_\pi^{-1} \sum_{y=1}^K (\int_0^1 \dot{g}_x(x, y)dx)^2 + \theta_c^{-1} \int_0^1 \sum_{y=1}^K \ddot{g}^2_{xx}(x, y)dx = \theta_\pi^{-1} \sum_{y=1}^K \beta_y^2 + \theta_c^{-1} \int_0^1 \sum_{y=1}^K \ddot{h}^2_{xx}(x, y)dx$. By keeping $\theta_\pi < \infty$, one shrinks β_y towards 0. When restricted to $J_\perp = \{b_j(y)\}_{j=1}^{K-1}$, the coefficients of $b_j(y)$ simply reparameterize $P(y|x) = P(Y = y|X = x) = p_y$, $\sum_{y=1}^K p_y = 1$, so the minimizer $\hat{g}$ of (1.2) uniquely exists as long as all K categories of the response are observed. If K is small, one may also choose to set $\theta_\pi = \infty$ to include $\{(x - .5)b_j(y)\}_{j=1}^{K-1}$ into $J_\perp$, in which case the minimizer of (1.2) uniquely exists whenever the maximum likelihood estimate exists for a linear logistic model of the form (after a reparameterization)

$$\begin{aligned} P(y|x) &= \frac{e^{\alpha_y + \beta_y x}}{1 + \sum_{z=1}^{K-1} e^{\alpha_z + \beta_z x}}, \quad y = 1, \ldots, K - 1, \\ P(K|x) &= \frac{1}{1 + \sum_{z=1}^{K-1} e^{\alpha_z + \beta_z x}}, \end{aligned} \tag{2.1}$$

where α_y, β_y, $y = 1, \cdots, K - 1$ are $2(K - 1)$ free parameters.

For $K = 2$, take $R_J = R^x_{c1} R^y_v$ and $J_\perp = \{b_1(y), (x - .5)b_1(y)\}$, where $R_v(1, 1) = b_1(1) = R_v(2, 2) = 1/2$ and $R_v(2, 1) = b_1(2) = R_v(1, 2) = -1/2$. It can be verified that the RK R_J generates an RKHS $\mathcal{H}_J = \{g(x, y) : g(x, 1) = -g(x, 2), \int_0^1 \ddot{g}^2_{xx}(x, 1)dx < \infty, \int_0^1 g(x, 1)dx = \int_0^1 \dot{g}_x(x, 1)dx = 0\}$ with a square norm $J(g) = \int_0^1 (\ddot{g}_{xx}(x, 1) - \ddot{g}_{xx}(x, 2))^2 dx/2 = 2\int_0^1 \ddot{g}^2_{xx}(x, 1)dx$. $J_\perp = \{g(x, y) : g(x, 1) = -g(x, 2) = \alpha + \beta x\}$. Now from $P(y|x) = e^{g(x,y)}/(e^{g(x,1)} + e^{g(x,2)})$, $y = 1, 2$ and $g(x, 1) = -g(x, 2)$, it is easy to check that $g(x, 1) = \log(P(1|x)/P(2|x))/2$. Hence, (1.2) in this setting simply reduces to the standard cubic spline logistic regression (cf. O'Sullivan, Yandell and Raynor (1986)).

Obviously, R_v is not the only nonnegative definite matrix which generates $\{\mathbf{1}\}^\perp$, and $\mathbf{1}^T\boldsymbol{g} = 0$ 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 a roughness functional $\sum_{y=1}^{K-1}(g(y+1) - g(y))^2$ in $\{\mathbf{1}\}^\perp$, which shrinks the differences between adjacent categories.

3. Computation of Estimates

We first derive a loss function for the assessment of estimation precision. Conditional on x, the symmetrized Kullback-Leibler distance between two conditional densities $e^g/\int_{\mathcal{Y}} e^g$ and $e^h/\int_{\mathcal{Y}} e^h$ is $\mathrm{SKL}(g, h|x) = \mu_g(g-h|x) - \mu_h(g-h|x)$, 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 $f(x)$, $\mathrm{SKL}(g, g_0) = \int_{\mathcal{X}} \mathrm{SKL}(g, g_0|x)f(x)$ 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 $\mathrm{SKL}(g - g_0)$ for g near g_0, where $v(h|x) = v(h, h|x)$.

The space $\mathcal{H}$ is in general infinite dimensional and $\hat{g}$ not computable. For the method to be of any use in practical estimation, one has to identify some appropriate finite dimensional function space in which estimates are to be calculated. A generic approach is to minimize (1.2) in $\mathcal{H}_n = J_\perp \oplus \mathrm{span}\{R_J(T_i, \cdot), i = 1, \cdots, n\}$, where R_J generates $\mathcal{H}_J$ and $T_i = (X_i, Y_i)$ are the observed data. When $g_0 \in J_\perp \oplus \mathcal{H}_J$, it can be shown that the minimizer $\hat{g}_n$ of (1.2) in $\mathcal{H}_n$ shares with $\hat{g}$ the same asymptotic convergence rates in $\mathrm{SKL}(g, g_0)$ and in $V(g - g_0)$ as $\lambda \to 0$ and $n \to \infty$ at certain rates, of course under suitable conditions, which provides a statistical justification for the choice of the adaptive space $\mathcal{H}_n$; see the Appendix. When $L(g|\mathrm{data})$ (cf. 1.1) depends on g only through evaluations $[T_i]g = g(T_i)$, such as in a logistic regression, then $\hat{g}_n = \hat{g}$; see, e.g., Wahba (1990) and O'Sullivan, Yandell and Raynor (1986). The existence theorem of Section 2 also applies to $\hat{g}_n$.

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

$$-\frac{1}{n}\mathbf{1}^T(Q\boldsymbol{c} + S\boldsymbol{d}) + \frac{1}{n}\sum_{i=1}^n \log\int_{\mathcal{Y}} \exp\{\boldsymbol{\xi}_i^T\boldsymbol{c} + \boldsymbol{\phi}_i^T\boldsymbol{d}\} + \frac{\lambda}{2}\boldsymbol{c}^TQ\boldsymbol{c} \tag{3.1}$$

with respect to $\boldsymbol{c}$ and $\boldsymbol{d}$, where Q is $n \times n$ with (i, j)th entry $\xi_i(T_j) = R_J(T_i, T_j) = J(\xi_i, \xi_j)$ and where $J(g, h)$ denotes the inner product associated with the square norm $J(g)$, S is $n \times M$ with (i,ν)th entry $\phi_\nu(T_i)$, $\boldsymbol{\xi}_i$ 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 $\tilde{g} = \boldsymbol{\xi}^T\tilde{\boldsymbol{c}} + \boldsymbol{\phi}^T\tilde{\boldsymbol{d}}$, the one-step Newton update for minimizing (3.1) can be shown to satisfy

$$\begin{pmatrix} V_{\xi,\xi} + \lambda Q & V_{\xi,\phi} \\ V_{\phi,\xi} & V_{\phi,\phi} \end{pmatrix} \begin{pmatrix} \boldsymbol{c} \\ \boldsymbol{d} \end{pmatrix} = \begin{pmatrix} Q\mathbf{1}/n - \boldsymbol{\mu}_\xi + V_{\xi,g} \\ S^T\mathbf{1}/n - \boldsymbol{\mu}_\phi + V_{\phi,g} \end{pmatrix}, \tag{3.2}$$

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 $V_{\phi,g} = V_{\tilde{g}}(\boldsymbol{\phi},\tilde{g})$.

Different smoothing parameters λ and θ_β (hidden in R_J) in (3.1) lead to different estimates, and ideally one would like to choose smoothing parameters which yield the best-performing estimate. A performance-oriented iteration scheme has been proposed, implemented, and evaluated for single smoothing parameter density estimation in Gu (1993a), which is designed to automatically land a hopefully well-performing estimate using information from the data. An implementation in multiple smoothing parameter problems has been further explored in Gu (1993b), which is directly applicable to conditional density estimation. Below we describe the general ideas behind the algorithms, and refer technical details to the aforementioned references.

Consider a single smoothing parameter problem with λ to be chosen but R_J fixed. As λ varies, one may perceive $\hat{g}_n$ as forming a "curve" in the function space $\mathcal{H}$, and the task is to locate a well-performing estimate on the "curve". With a varying λ, the one-step Newton updates of (3.2) from $\tilde{g}$ may be perceived as forming a "line" in $\mathcal{H}$, and our strategy is to move onto the best-performing estimate on the "line" in each iteration. Assuming that a means does exist for the comparison of the performances of the estimates on such "lines", and that the iteration does converge to a fixed point (λ_*, g_*), it is easy to see that g_* is indeed on the "curve" with $\lambda = \lambda_*$, and there is no estimate on the "line" of one-step Newton updates from g_* that performs better than g_*. We seek such a g_* as our automatic estimate. The same scenario holds for multiple smoothing parameters, with the "curve" replaced by a "surface" and the "lines" replaced by "planes".

In conditional density estimation, we shall use $\mathrm{SKL}(g,g_0)$ or $V(g-g_0)$, which are proxies of each other, to measure the performance of the estimate, and for the observed data, we shall replace the density $f(x)$ of X appearing in $\mathrm{SKL}(g,g_0)$ and $V(g-g_0)$ by the empirical distribution of X_i, as in the definitions of $\mu_g(h)$ and $V_g(h,f)$ above. For an estimate g on the "line" of one-step Newton updates from $\tilde{g}$, there exists a proxy of $\mathrm{SKL}(g,g_0)$ or $V(g-g_0)$ consisting of three terms: a term depending only on g_0 and $\tilde{g}$, which can be ignored for comparative purposes, a term depending only on g and $\tilde{g}$, which can be computed, and a cross term of the form $\mu_{g_0}(g)$, whose components can be estimated by the corresponding sample means or some cross-validated versions of sample means. Using such estimated proxies of the performance measure on the "lines" to compare estimates, one may drive the performance-oriented iteration to land a hopefully near optimal estimate on the "curve".

With a careful implementation, such an iteration scheme rarely diverges in our empirical studies under various settings. The algorithm takes $J_\perp$, R_J, and the data as inputs, and returns an estimate with automatic smoothing parameters, if it converges.

4. Applications

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

4.1. 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 $\int_\mathcal{X} g_x = \int_\mathcal{Y} g_y = \int_\mathcal{X} g_{x,y} = \int_\mathcal{Y} g_{x,y} = 0$, 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 trimmed 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 4.1, 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 the 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 standard approach to regression with known breaks is to add jumps at breaks, using 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 the tensor products 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 trimmed 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 sketched in the right frame of Figure 4.1 in a manner similar to the left frame. Apart from the downward shift, the coins seem to get thicker at a steady pace as time goes on.

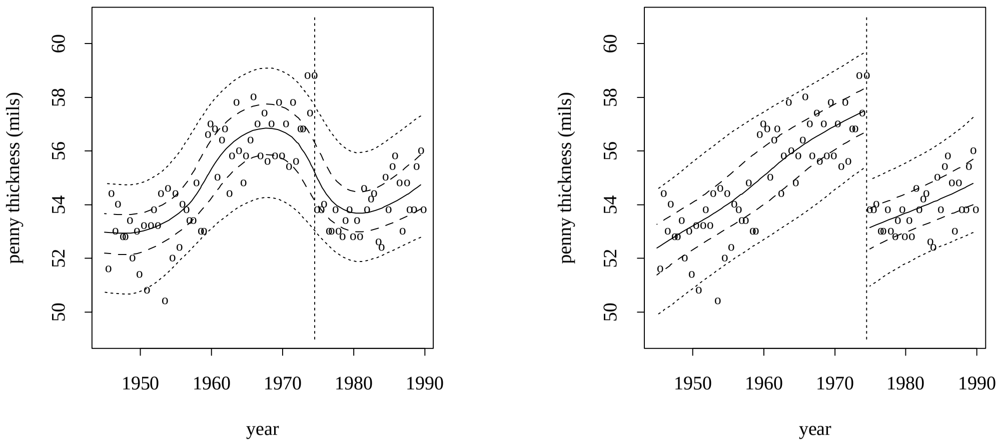

Figure 4.1. Penny Thickness Data. Left: continuous model; right: model with break. Solid lines are conditional medians, dashed lines quartiles, and dotted lines 5th and 95th percentiles. Circles are data and vertical dotted line position of the break.

Denote by $f_1(y|x)$ the estimate in the left frame of Figure 4.1 and by $f_2(y|x)$ that in the right. $f_2(y|x)$ seems to fit the data slightly better, in the sense that the log likelihood $\sum_{i=1}^{90}\log f_2(Y_i|X_i) = 85.21$ comes out slightly bigger than $\sum_{i=1}^{90}\log f_1(Y_i|X_i) = 83.93$. The two configurations took about 80 and 90 cpu minutes to compute, respectively, on an IBM-RS6000.

4.2. 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, 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 a logistic regression and the latter is a 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 $1 + R_v^{x_1}$ with RK's on the (product) continuous domain, where $R_v^{x_1}$ on $\{1, 2\}^2$ can be written as a 2×2 matrix $(I - \mathbf{1}\mathbf{1}^T/2)$. Since a simple constant shift may suffice in this context, I chose to cut off all but one product term which involves $R_v^{x_1}$, that of $R_v^{x_1}$ with the 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 RK on the covariate domain is thus taken as $1 + R_v^{x_1} + R_{\pi 1}^{x_2} + R_{c1}^{x_2} + R_{\pi 1}^{x_3} + R_{c1}^{x_3} + R_{\pi 1}^{x_2} R_{\pi 1}^{x_3} + R_{c1}^{x_2} R_{\pi 1}^{x_3} + R_{\pi 1}^{x_2} R_{c1}^{x_3} + R_{c1}^{x_2} R_{c1}^{x_3}$, with each term subject to a scaling. On the y axis one may simply use R_v^y as the RK, which can be written as a 2×2 matrix (for logistic regression) or a 5×5 matrix (for multinomial regression). Taking the tensor product of the RKHS's 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 $J_\perp = \{R_v^y(y, j), R_v^{x_1}(x_1, 1) R_v^y(y, j)\}_{j=1}^{K-1}$ with dimension $2(K-1)$, where K is 2 or 5, the number of diagnostic categories. Note that $R_v^{x_1}$ and R_v^y are different when $K = 5$.

For multinomial regression, the performance oriented iteration effectively trimmed all but two penalized terms, leaving only $\theta_{\pi,0}, \theta_{0,\pi} > 0$. This leads to a parametric logistic model of the form $g(x_1, x_2, x_3, y) = \alpha_{x_1,y} + \beta_y(x_2 - .5) + \gamma_y(x_3 - .5)$ with $\sum_y \alpha_{1,y} = \sum_y \alpha_{2,y} = \sum_y \beta_y = \sum_y \gamma_y = 0$ (cf. 2.1). There are however significant differences between our nonparametric-turned-parametric estimate and the maximum likelihood estimate directly based on such a parametric model. One difference is that we have allowed more flexibility in the procedure, but our performance estimates implied that the extra flexibility did not seem to be advantageous for the given data. Another difference is that the parameters β_y and γ_y have been shrunk towards zero, with the amount of shrinking tuned automatically according to our performance estimates.

For logistic regression, the performance oriented iteration with the "same" eight penalized terms enrountered numerical difficulty as $\theta_{0,\pi}$ was moving towards ∞. This indicates that the one dimensional space $\{(x_3 - .5) R_v(y, 1)\}$ generated by the RK $R^{x_3}_{\pi 1}R^{y}_{v}$ should have been put into $J_\perp$ instead of being penalized. The iteration with $(x_3 - .5)R_v(y,1)$ added to $J_\perp$ and $\theta_{0,\pi}R^{x_3}_{\pi 1}R^{y}_{v}$ removed from R_J converged without incidence, effectively leaving only $\theta_{\pi,0}, \theta_{\pi,c} > 0$.

One reason that none of the "nonparametric" terms survived the automatic trimming for $K = 5$ but one survived for $K = 2$ might be due to the fact that $J_\perp$ has 8 dimensions for $K = 5$ whereas the augmented $J_\perp$ only has 3 dimensions for $K = 2$. Also, the three penalized "parametric" terms for $K = 5$, of which two survived, each have 4 dimensions, but the two penalized terms for $K = 2$ each have only 1 dimension.

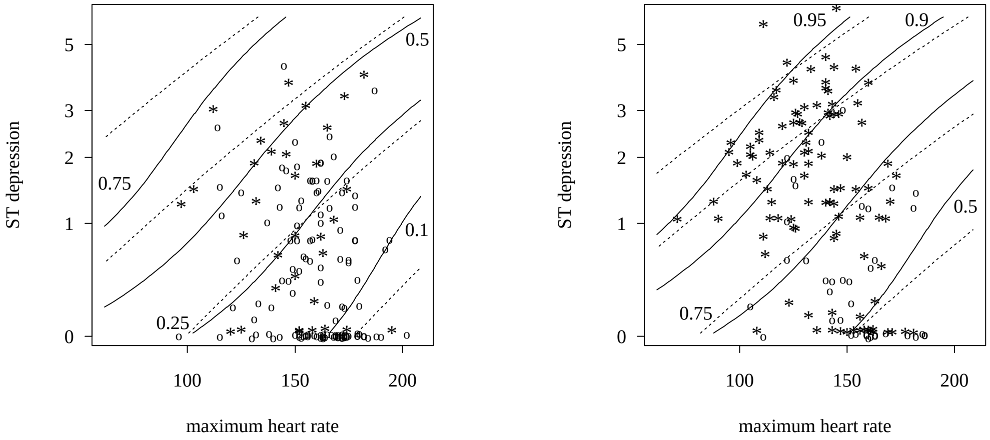

Figure 4.2. Heart Disease Data. Left: disease rates with symptomatic chest pain; right: disease rates with asymptomatic chest pain. Solid lines are estimates from logistic model and dotted lines estimates from multinomial model. Circles are healthy patients and stars disease patients.

The estimated disease rate, the probability that a patient has angiographic disease, is shown in Figure 4.2, where the solid lines are from the logistic model and the dotted lines are 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 areas. The logistic and multinomial estimates took about 138 and 246 cpu minutes to compute, respectively, on an IBM-RS6000.

Following a suggestion by the Associate Editor, we also calculated the parametric maximum likelihood estimates of form $g(x_1,x_2,x_3,y) = \alpha_{x_1,y} + \beta_y(x_2 - .5) + \gamma_y(x_3 - .5) + \delta_y(x_2 - .5)(x_3 - .5)$, $\sum_y \alpha_{1,y} = \sum_y \alpha_{2,y} = \sum_y \beta_y = \sum_y \gamma_y = \sum_y \delta_y = 0$, for $K = 2$ and of form $g(x_1,x_2,x_3,y) = \alpha_{x_1,y} + \beta_y(x_2 - .5) + \gamma_y(x_3 - .5)$,

$\sum_y \alpha_{1,y} = \sum_y \alpha_{2,y} = \sum_y \beta_y = \sum_y \gamma_y = 0$, for $K = 5$, without shrinking. The estimated disease rate is shown in Figure 4.3, where the dashed lines are from the parametric logistic model, the dotted lines are aggregated from the parametric multinomial model, and the solid lines of Figure 4.2 are superimposed for comparison. The 5 degrees-of-freedom parametric interactive logistic estimate with a log likelihood $\sum_{i=1}^{303} \log f(Y_i|X_i) = -149.29$ does not seem to fit the data as well, as compared to the spline logistic estimate with a log likelihood $\sum_{i=1}^{303} \log f(Y_i|X_i) = -142.77$. It is interesting to observe that the estimate aggregated from the 16 degrees-of-freedom parametric multinomial fit comes out extremely close to the spline logistic estimate.

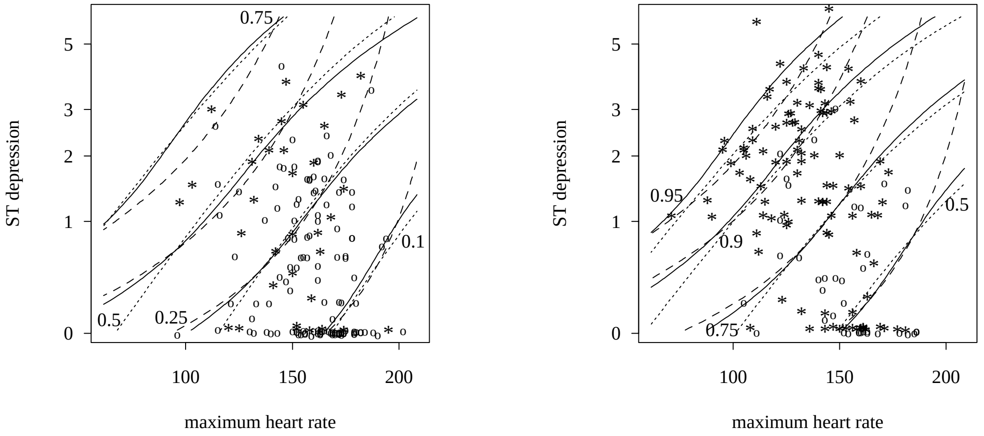

Figure 4.3. Heart Disease Data. Left: disease rates with symptomatic chest pain; right: disease rates with asymptomatic chest pain. Solid lines are spline estimates from logistic model, dashed lines parametric estimates from logistic model, and dotted lines parametric estimates from multinomial model. Circles are healthy patients and stars disease patients.

5. 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 facilitates research in this direction.

When $K = 2$ in Example 2.3, it is seen that the estimation via (1.2) using R_v on the y axis is equivalent to the standard penalized likelihood logistic regression, 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 (1992) designed for a computation scheme in which the penalized likelihood problem is solved via a sequence of penalized weighted least squares problems. Further empirical study is needed to compare the two methods for smoothing parameter selection in logistic regression.

Acknowledgements

The author thanks the referees for constructive comments. This research was supported by a Summer Faculty Grant from Purdue Research Foundation and grants from National Science Foundation.

Appendix: Asymptotic Theory

We describe an asymptotic theory, which extends theoretical support to the proposed method. The theory is parallel to that in Gu and Qiu (1993), where technical details are to be found. We shall assume that $g_0 \in \mathcal{H}$, and that the maximum likelihood estimate exists in $J_\perp$ so $\hat{g}$ and $\hat{g}_n$ exist. The theory concerns the asymptotic convergence rates of $\hat{g}$ and $\hat{g}_n$.

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 $J(g)$. 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 $0 < \beta_1 < \beta_2 \le \infty$.

For Examples 2.1 and 2.2, it can be shown that 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 with a cubic spline on the $\mathcal{X}$ domain, $r = 4$. See Gu (1993c) for relevant technical details.

Define $Q_\lambda(g) = -(1/n)\sum_{i=1}^n \{g(X_i, Y_i) - \mu_{g_0}(g|X_i)\} + (1/2)V(g - g_0) + (\lambda/2)J(g)$, a quadratic approximation of (1.2) at g_0. Let g_1 be the minimizer of $Q_\lambda(g)$ in $\mathcal{H}$, which exists similar to $\hat{g}$. The convergence rates of $\hat{g}$ and $\hat{g}_n$ are based on that of g_1.

Theorem A.1. *Under A.1 and A.2, as $\lambda \to 0$ and $n\lambda^{1/r} \to \infty$, $(V + \lambda J)(g_1 - g_0) = O_p(n^{-1}\lambda^{-1/r} + \lambda)$.*

We need a few more technical assumptions for the rest of the theory.

Assumption A.3. For g in a convex set B_0 around g_0 containing $\hat{g}$, $\hat{g}_n$ and g_1,

$$\exists c_1, c_2 \in (0, \infty) \text{ such that } c_1 v(h|x) \le v_g(h|x) \le c_2 v(h|x),\ \forall h \in \mathcal{H}, \forall x \in \mathcal{X}.$$

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

$$\int_{\mathcal{X}} [v(\phi_\nu \phi_\mu|x) + \{\mu_{g_0}(\phi_\nu\phi_\mu|x) - \int_{\mathcal{X}} \mu_{g_0}(\phi_\nu\phi_\mu|x) f(x)\}^2] f(x) \le c_4,\ \forall \nu, \mu.$$

Note that $v_g(f - h|x) = \mu_g((f-h)^2|x) - \mu_g^2(f - h)$ is a form of (weighted) mean square error between f and h with e^g as the weight; thus A.3 is simply assuming that minor changes in the weight function do not change the order of magnitude of the mean square error, which appears mild. 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 that $v(g - h|x)$ and $\mathrm{SKL}(g, h|x)$, and hence $V(g - h)$ and $\mathrm{SKL}(g, h)$, are of the same order of magnitude for $g, h \in B_0$. A.4 and A.5 are essentially moment conditions on the Fourier basis $\phi_\nu(X, Y)$: A uniform bound on the fourth moments of $\phi_\nu(X, Y)$ is sufficient for A.4 and A.5 to hold. Although explicit forms are in general not available, $\phi_\nu(X, Y)$ will grow in frequency as $\nu \to \infty$ as its definition suggests, but it is unlikely to grow indefinitely in magnitude since

$V(\phi_\nu) = 1$. Hence, such moment conditions seem highly plausible. A.3 – A.5 may not be easily verifiable from more primitive conditions, however.

The convergence rates of $\hat{g}$ and $\hat{g}_n$ are summarized in the following theorems, with the latter supporting our use of $\hat{g}_n$ in practice.

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

Theorem A.3. *Modify* A.3 *to also include* 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 proofs of Theorems A.1-A.3 are straightforward adaptations of those in Gu and Qiu (1993), and hence are omitted.

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

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

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

Gu, C. (1993c). Penalized likelihood hazard estimation: A general procedure with asymptotic theory. Technical Report 91-58 (Rev.), 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. (1993). Smoothing spline ANOVA with component-wise Bayesian confidence intervals. J. Comput. Graph. Statist. **2**, 97 – 117.

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. S. and Raynor, W. J. (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. John 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 hypotheses, conditional independence graphs and graphical chain models. J. Roy. Statist. Soc. Ser.B **52**, 21 – 50.

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

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

(Received July 1993; accepted November 1994)