Statistica Sinica **13**(2003), 811-826

PENALIZED LIKELIHOOD DENSITY ESTIMATION: DIRECT CROSS-VALIDATION AND SCALABLE APPROXIMATION

Chong Gu and Jingyuan Wang

Purdue University

Abstract: For smoothing parameter selection in penalized likelihood density estimation, a direct cross-validation strategy is illustrated. The strategy is as effective as the indirect cross-validation developed earlier but is much easier to implement in multivariate settings. Also studied is the practical implementation of certain low-dimensional approximations of the estimate, with the dimension of the model space selected to achieve both asymptotic efficiency and numerical scalability. The greatly reduced computational burden allows the routine use of the technique for the analysis of large data sets. Related practical issues concerning multivariate numerical integration are also briefly addressed.

Key words and phrases: Cross-validation, Kullback-Leibler loss, penalized likelihood, smoothing parameter.

1. Introduction

Consider the estimation of a probability density $f(x)$ on a domain $\mathcal{X}$ based on independent samples X_i, $i = 1, \cdots, n$. In classical parametric estimation, some parametric form is assumed of $f(x)$ and the unknown parameters are commonly estimated by maximum likelihood. When adequate parametric models are not available, various nonparametric methods can be called to service. This article concerns one such nonparametric method, the penalized likelihood method, which was pioneered by Good and Gaskins (1971) and further developed by Silverman (1982), O'Sullivan (1988), Gu and Qiu (1993) and Gu (1993).

Assume a bounded domain $\mathcal{X}$ so that the uniform density is proper. Write $f(x) = e^{\eta(x)}/\int_{\mathcal{X}} e^{\eta(x)}$, the logistic density transform (Leonard (1978)). The estimation of $f(x)$ can be conducted through the minimization of a penalized likelihood functional,

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

where the first term is the minus log likelihood, $J(\eta)$ is a quadratic roughness functional, and λ, known as the smoothing parameter, controls the tradeoff between the goodness-of-fit and the smoothness of the estimate. To make the logistic density transform one-to-one, Gu and Qiu (1993) proposed to impose a side condition on η; two examples of side conditions are $\eta(x_0) = 0$ for some $x_0 \in \mathcal{X}$ and $\int_{\mathcal{X}} \eta(x) = 0$. With $\lambda = \infty$, one enforces a parametric model in the null space of $J(\eta)$, $\mathcal{N}_J = \{\eta : J(\eta) = 0\}$, and when $\lambda = 0$, one obtains the nonparametric maximum likelihood estimate that corresponds to the empirical distribution. The optimal λ, in a sense to be made clear later, is somewhere in between.

A simple example of (1) on a one-dimensional domain $\mathcal{X} = [0, 1]$ is the cubic spline density estimate with $J(\eta) = \int_0^1 \ddot{\eta}^2(x)dx$ and $\mathcal{N}_J = \text{span}\{x + C\}$, where the constant C is determined by the side condition imposed on η.

When $\mathcal{X}$ is a product domain, certain ANOVA decompositions can be built into penalized likelihood estimation using tensor product splines. For example, with $X = (U, V) \in \mathcal{X} = \mathcal{U} \times \mathcal{V}$, one may construct

$$\eta(x) = \eta(u, v) = \eta_\emptyset + \eta_u(u) + \eta_v(v) + \eta_{u,v}(u, v), \tag{2}$$

where $\eta_\emptyset$ is the constant, η_u and η_v are the main effects, and $\eta_{u,v}$ is the u-v interaction. The identifiability of the decomposition can be assured through certain side conditions imposed on η_u, η_v and $\eta_{u,v}$, and the constant $\eta_\emptyset$ is to be eliminated for a one-to-one transform $f(u, v) = e^{\eta_u+\eta_v+\eta_{u,v}} / \int_{\mathcal{X}} e^{\eta_u+\eta_v+\eta_{u,v}}$. An additive model $\eta = \eta_u + \eta_v$ characterizes the independence of U and V. For $\mathcal{X}$ with multiple marginals, selective term elimination in such ANOVA decompositions yields various conditional independence structures. This provides a means to the nonparametric estimation of certain graphical models; see, e.g., Whittaker (1990) for an introduction to graphical models and their parametric estimation. Technical details concerning the construction of tensor product splines can be found in, e.g., Wahba (1990) and Gu (2002).

As is well known, the key to successful nonparametric estimation is to strike a proper balance between "bias" and "variance." In penalized likelihood estimation, such balance is to be achieved through the proper selection of λ. Too small a λ yields very rough estimates, or too much "variance," and too large a λ allows little flexibility beyond the null space $\mathcal{N}_J$ of $J(\eta)$, leaving more "bias" in the estimate. In multivariate problems with the aforementioned ANOVA decompositions, the roughness functional decomposes accordingly to a sum of component roughness functionals, say $J(\eta) = \theta_u^{-1} J_u(\eta_u) + \theta_v^{-1} J_v(\eta_v) + \theta_{u,v}^{-1} J_{u,v}(\eta_{u,v})$ for the decomposition in (2), where the θ's, an extra set of smoothing parameters to be selected, determine the relative weights of component smoothness.

The seminal work of Craven and Wahba (1979) on generalized cross-validation laid the foundation for smoothing parameter selection in penalized likelihood estimation, and remains the method of choice for Gaussian regression. O'Sullivan (1988) adapted a certain cross-validation score in the kernel method literature to calculate the density estimate of Silverman (1982). Gu (1993) developed an indirect cross-validation approach to the selection of λ in (1), but the numerical efficiency of its multiple smoothing parameter implementation was less than ideal (Gu (1998)). We demonstrate in this article a direct cross-validation strategy for use with (1), more effective and numerically more efficient than the indirect cross-validation of Gu (1993).

The minimizer of (1) in an infinite dimensional function space is generally not computable. Based on an asymptotic analysis of Gu and Qiu (1993), to be reviewed shortly, Gu (1993) considered a certain finite dimensional approximation that requires $O(n^3)$ flops (floating point operations) to compute. A trivial refinement of the asymptotic analysis yields approximations that are much faster to compute yet maintain the same asymptotic efficiency. Various issues concerning the practical implementation of such faster approximations will also be addressed in this article.

The rest of the article is organized as follows. In §2, pertinent technical details concerning (1) are reviewed and the numerical problem is formulated; the asymptotic analysis mentioned above is discussed but the details are relegated to the appendix. In §3, the direct cross-validation strategy is outlined and related computation is discussed. Through simulation studies, §4 demonstrates the empirical performance of the cross-validation score and develops strategies for the practical calculation of the estimates. Numerical integration is needed in the implementation of the method, of which a few practical issues are discussed in §5. An illustrative data example is given in §6. A few remarks are collected in §7 to conclude the article. Part of this work has been excerpted by the first author for use in a monograph (Gu (2002)); notes are added in the text and relevant details are omitted to reflect the overlap.

2. Formulation and Notation

The functional (1) is defined in a Hilbert space $\mathcal{H} \subseteq \{\eta : J(\eta) < \infty\}$, in which $J(\eta)$ is a square (semi) norm with a finite dimensional null space $\mathcal{N}_J = \mathcal{H} \cap \{\eta : J(\eta) = 0\}$. A Hilbert space has a metric and a geometry, which facilitates the analysis and the computation of the estimate, and a finite dimensional $\mathcal{N}_J$ prevents interpolation (i.e., the empirical distribution). One needs the evaluation functional $[x]\eta = \eta(x)$ to be continuous in $\mathcal{H}$ so that the first term of (1) is continuous, and the members of $\mathcal{H}$ have to comply with a side condition mentioned earlier to make the second term strictly convex.

A Hilbert space in which evaluation is continuous is a reproducing kernel Hilbert space with a reproducing kernel $R(\cdot,\cdot)$, a non-negative definite bivariate function on $\mathcal{X}$ such that $R(x,\cdot) = R(\cdot,x) \in \mathcal{H}$, $\forall x \in \mathcal{X}$, and $\langle R(x,\cdot), \eta(\cdot)\rangle = \eta(x)$ (the reproducing property), $\forall \eta \in \mathcal{H}$, where $\langle\cdot,\cdot\rangle$ is the inner product in $\mathcal{H}$; see

Aronszajn (1950). For the discussion of this article, one needs a basis of $\mathcal{N}_J$ and a reproducing kernel $R_J(x,y)$ in $\mathcal{H}\ominus\mathcal{N}_J$.

As a concrete example, consider again the cubic spline on $\mathcal{X}=[0,1]$ with $J(\eta)=\int_0^1 \ddot{\eta}^2(x)dx$. Different side conditions lead to different R_J and $\mathcal{N}_J$, but the density estimate after the transform $f(x)=e^{\eta(x)}/\int_0^1 e^{\eta(x)}$ remains the same. If one specifies $\int_0^1 \eta(x)=0$, then $\mathcal{N}_J=\{(\cdot-0.5)\}$ and $R_J(x,y)=k_2(x)k_2(y)-k_4(|x-y|)$, where $k_2=(k_1^2-1/12)/2$, $k_4=(k_1^4-k_1^2/2+7/240)/24$, with $k_1=(\cdot-0.5)$. If one specifies $\eta(0)=0$, then $\mathcal{N}_J=\{(\cdot)\}$ and $R_J(x,y)=\int_0^1(x-u)_+(y-u)_+du$, where $(\cdot)_+$ is the positive part of $(\cdot)$. Further details and more examples can be found in Gu and Qiu (1993) and Gu (1998).

The space $\mathcal{H}$ is usually infinite dimensional, and the minimizer of (1) in $\mathcal{H}$ is in general not computable. To circumvent the problem, Gu and Qiu (1993) proposed to use the minimizer of (1) in an adaptive finite dimensional space $\mathcal{H}_n=\mathcal{N}_J\oplus\{R_J(X_i,\cdot),i=1,\ldots,n\}$, which was the estimate calculated in Gu (1993, 1998). Under mild conditions, the minimizer of (1) in $\mathcal{H}_n$ was shown by Gu and Qiu (1993) to share the same asymptotic convergence rates as the minimizer in $\mathcal{H}$. A careful look at the theory reveals that one could actually achieve the same convergence rates in a space $\mathcal{H}_q=\mathcal{N}_J\oplus\{R_J(Z_j,\cdot),j=1,\ldots,q\}$ with $q\asymp n^{2/(pr+1)+\epsilon}$ for some $p\in[1,2]$, $r>1$, $\forall\epsilon>0$, where Z_j, $j=1,\ldots,q$ is a random subset of X_i; details are to be found in the appendix. The constant p depends on how smooth the "true" η is: for the cubic spline on $\mathcal{X}=[0,1]$, $p=1$ if $\ddot{\eta}^2$ is "barely" integrable, and $p=2$ if $\eta^{(4)}$ is square integrable.

The computation of the minimizer in $\mathcal{H}_q$ is of the order $O(nq^2)+O(dq^2)$, where d is the size of the quadrature for integration, representing significant savings over the $O(n^3)+O(dn^2)$ needed to work with $\mathcal{H}_n$; see §3 for details. For the cubic spline example given above, $r=4$, so $q=O(n^{2/(4p+1)+\epsilon})$. For "supersmooth" η with $p=2$, one could make do with only $O(n^{13/9+\epsilon})+O(dn^{4/9+\epsilon})$ computation, and for rougher η that nevertheless satisfies $J(\eta)<\infty$, the computational burden is no worse than $O(n^{9/5+\epsilon})+O(dn^{4/5+\epsilon})$.

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

$$\eta=\sum_{\nu=1}^m d_\nu\phi_\nu+\sum_{j=1}^q c_j\xi_j=\boldsymbol{\phi}^T\boldsymbol{d}+\boldsymbol{\xi}^T\boldsymbol{c}, \tag{3}$$

where $\boldsymbol{\phi}$ and $\boldsymbol{\xi}$ are vectors of functions and $\boldsymbol{d}$ and $\boldsymbol{c}$ are vectors of coefficients. Substituting (3) into (1), noting that $J(\eta)=\langle\sum_{j=1}^q c_j\xi_j\,,\sum_{k=1}^q c_k\xi_k\rangle=\sum_{j=1}^q\sum_{k=1}^q c_jc_kR_J(Z_j,Z_k)$, one calculates the minimizer η_λ of (1) in $\mathcal{H}_q$ by minimizing

$$A_\lambda(\boldsymbol{d},\boldsymbol{c})=-\frac{1}{n}\mathbf{1}^T(S\boldsymbol{d}+R\boldsymbol{c})+\log\int_{\mathcal{X}}\exp(\boldsymbol{\phi}^T\boldsymbol{d}+\boldsymbol{\xi}^T\boldsymbol{c})+\frac{\lambda}{2}\boldsymbol{c}^TQ\boldsymbol{c}$$

with respect to $\boldsymbol{d}$ and $\boldsymbol{c}$, where S is $n \times m$ with the (i,ν)th entry $\phi_\nu(X_i)$, R is $n \times q$ with the (i,j)th entry $\xi_j(X_i) = R_J(X_i, Z_j)$, and Q is $q \times q$ with the (j,k)th entry $\xi_j(Z_k) = R_J(Z_j, Z_k)$.

The convex function $A_\lambda(\boldsymbol{d},\boldsymbol{c})$ may be minimized via Newton iteration. Write $\mu_\eta(g) = \int_{\mathcal{X}} g e^\eta / \int_{\mathcal{X}} e^\eta$ and $V_\eta(g,h) = \mu_\eta(gh) - \mu_\eta(g)\mu_\eta(h)$. Let $\tilde{\eta} = \boldsymbol{\phi}^T \tilde{\boldsymbol{d}} + \boldsymbol{\xi}^T \tilde{\boldsymbol{c}}$ be the current iterate of η. Straightforward calculation yields the updating equation

$$\begin{pmatrix} V_{\phi,\phi} & V_{\phi,\xi} \\ V_{\xi,\phi} & V_{\xi,\xi} + \lambda Q \end{pmatrix} \begin{pmatrix} \boldsymbol{d} \\ \boldsymbol{c} \end{pmatrix} = \begin{pmatrix} S^T \mathbf{1}/n - \mu_\phi + V_{\phi,\eta} \\ R^T \mathbf{1}/n - \mu_\xi + V_{\xi,\eta} \end{pmatrix}, \tag{4}$$

where $V_{\phi,\phi} = V_{\tilde{\eta}}(\boldsymbol{\phi}, \boldsymbol{\phi}^T)$, $V_{\phi,\xi} = V_{\tilde{\eta}}(\boldsymbol{\phi}, \boldsymbol{\xi}^T)$, $V_{\xi,\xi} = V_{\tilde{\eta}}(\boldsymbol{\xi}, \boldsymbol{\xi}^T)$, $\mu_\phi = \mu_{\tilde{\eta}}(\boldsymbol{\phi})$, $\mu_\xi = \mu_{\tilde{\eta}}(\boldsymbol{\xi})$, $V_{\phi,\eta} = V_{\tilde{\eta}}(\boldsymbol{\phi}, \tilde{\eta})$, and $V_{\xi,\eta} = V_{\tilde{\eta}}(\boldsymbol{\xi}, \tilde{\eta})$; see, e.g., Gu (1993) and Gu (2002, §6.1).

Equation (4) forms the basis for computation. For a multiple term penalty $J(\eta) = \sum_\beta \theta_\beta^{-1} J_\beta(\eta)$ such as the ones associated with the tensor product splines, the reproducing kernel of $\mathcal{H}_J$ is of the form $R_J = \sum_\beta \theta_\beta R_\beta$, so beside the smoothing parameter λ appearing explicitly in (4), one also has the θ's hidden in the entries involving $\boldsymbol{\xi}$.

3. Cross-Validation and Computation

We now outline the direct cross-validation strategy and the associated computation. Much of this material has been presented by the first author in Gu (2002, §6.3, §6.4 in a slightly different notation, so most of the derivation is omitted here to minimize the overlap.

To measure the proximity of the estimate $f_\lambda = e^{\eta_\lambda} / \int_{\mathcal{X}} e^{\eta_\lambda}$ to the true density $f = e^\eta / \int_{\mathcal{X}} e^\eta$, we use the Kullback-Leibler distance $\mathrm{KL}(\eta, \eta_\lambda) = E_f \log(f/f_\lambda) = \mu_\eta(\eta - \eta_\lambda) - \log \int_{\mathcal{X}} e^\eta + \log \int_{\mathcal{X}} e^{\eta_\lambda}$. The smoothing parameters that minimize $\mathrm{KL}(\eta, \eta_\lambda)$ are considered the optimal ones.

Dropping terms that do not involve η_λ, one obtains the relative Kullback-Leibler distance, $\mathrm{RKL}(\eta, \eta_\lambda) = \log \int_{\mathcal{X}} e^{\eta_\lambda} - \mu_\eta(\eta_\lambda)$; the term $\mu_\eta(\eta_\lambda)$ involving the unknown density will have to be estimated. A naive estimate of $\mu_\eta(\eta_\lambda)$ is the sample mean $n^{-1} \sum_{i=1}^n \eta_\lambda(X_i)$, but it is biased because the samples X_i contribute to the estimate η_λ; the resulting estimate of $\mathrm{RKL}(\eta, \eta_\lambda)$ is simply minus the log likelihood, clearly favoring $\lambda = 0$. Standard cross-validation suggests an estimate $n^{-1} \sum_{i=1}^n \eta_\lambda^{[i]}(X_i)$, where $\eta_\lambda^{[i]}$, the minimizer of the delete-one version of (1), is however expensive to compute. For an analytically tractable approximation of $\eta_\lambda^{[i]}$, we consider the quadratic approximation of (1) at $\tilde{\eta} = \eta_\lambda$, and compute the minimizer $\eta_{\lambda,\tilde{\eta}}^{[i]}$ of the delete-one version thereof. Estimating $\mu_\eta(\eta_\lambda)$ in $\mathrm{RKL}(\eta, \eta_\lambda)$ by $n^{-1} \sum_{i=1}^n \eta_{\lambda,\tilde{\eta}}^{[i]}(X_i)$, one has, for $\alpha = 1$,

$$-\frac{1}{n} \sum_{i=1}^n \left\{ \eta_\lambda(X_i) - \log \int_{\mathcal{X}} e^{\eta_\lambda} \right\} + \alpha \frac{\mathrm{tr}(P_{\mathbf{1}}^\perp \tilde{R} H^{-1} \tilde{R}^T P_{\mathbf{1}}^\perp)}{n(n-1)}, \tag{5}$$

where H is the left-hand-side matrix in (4), $\tilde{R} = (S, R)$, and $P_{\mathbf{1}}^{\perp} = I - \mathbf{1}\mathbf{1}^T/n$ of size $n \times n$. Detailed derivation, with a slightly different notation for $q = n$, can be found in Gu (2002, §6.3).

As is shown in §4, the practical performance of the cross-validation score (5) with $\alpha = 1$ is generally adequate but, as is typical with cross-validation techniques, the method may severely undersmooth up to about 10% of the replicates in simulation studies. To circumvent the problem, a simple modification of the score proves to be remarkably effective. The score is seen to have a rather simple structure, with minus the log likelihood monotonically decreasing as λ decreases while the trace term moves in the opposite direction. To force smoother estimates, one may simply set $\alpha > 1$ in (5). Simulation studies suggest that an α around 1.4 would be most effective, curbing undersmoothing on "bad" replicates while sacrificing minimal performance degradation on "good" ones; details are in §4.

Fixing smoothing parameters, the computation involves Newton iteration via (4) and the evaluation of the cross-validation score (5). This can be accomplished by a Cholesky decomposition of H followed by forward and backward substitutions; see Gu (2002, §6.4 for details. To select smoothing parameters by cross-validation, standard optimization tools such as those developed in Dennis and Schnabel (1996) can be used. These employ a certain quasi-Newton approach with numerical derivatives.

The one-time formation of $\tilde{R}^T\mathbf{1}$ takes $O(nq)$ flops. The formation of the other entries in (4) takes $O(dq^2)$ flops, where d is the size of the quadrature for integration. The Cholesky decomposition is of the order $O(q^3)$ and the back and forward substitutions are of the order $O(q^2)$. For the cross-validation score, the log likelihood part takes $O(nq)$ flops and the trace part takes $O(nq^2)$ flops (n forward substitutions). All in all, the computation is of the order $O(nq^2)+O(dq^2)$.

With the indirect cross-validation of Gu (1993), $(\tilde{\eta}, \lambda)$ are jointly updated through (4): $\tilde{\eta}$ is updated by the minimizer $\eta_{\lambda,\tilde{\eta}}$ of the quadratic approximation of (1) at $\tilde{\eta}$, for λ minimizing a $\tilde{\eta}$-specific cross-validation score that tracks $\mathrm{KL}(\eta, \eta_{\lambda,\tilde{\eta}})$ as a function of λ. Roughly speaking, the indirect method nests smoothing parameter selection under Newton iteration, while the direct method outlined here nests Newton iteration under smoothing parameter selection. Since smoothing parameter selection by cross-validation is much more involved than Newton iteration, especially for multiple smoothing parameters, the direct method is numerically more efficient. Also, a fixed point always exists for Newton iteration with fixed smoothing parameters, but the joint updating of $(\tilde{\eta}, \lambda)$ may never settle down.

4. Simulation Studies

In this section, we demonstrate the empirical performance of cross-validation and explore empirical rules for the choice of q through simulation studies.

4.1. Empirical performance of cross-validation

First consider a univariate test density

$$f_1(x) \propto \Big\{\frac{1}{3}e^{-50(x-0.3)^2} + \frac{2}{3}e^{-50(x-0.7)^2}\Big\}I_{[0<x<1]}, \tag{6}$$

which is a mixture of $N(0.3, 0.01)$ and $N(0.7, 0.01)$ truncated to $[0, 1]$. Samples of size $n = 100$ were drawn from the density, and cubic splines were used with $J(\eta) = \int_0^1 \ddot{\eta}^2(x)dx$, $\mathcal{N}_J = \{(\cdot - 0.5)\}$, $R_J(x, y) = k_2(x)k_2(y) - k_4(|x - y|)$, and $q = n$; see §2 for the notation. Estimates η_λ were calculated on the grid $\log_{10}\lambda = (-7)(0.1)(-3)$ for 100 replicates. Recorded for each of the estimates were the Kullback-Leibler distance $\mathrm{KL}(\eta, \eta_\lambda)$, the minus log likelihood $n^{-1}\sum_{i=1}^n\{-\eta_\lambda(X_i)+\log\int_{\mathcal{X}} e^{\eta_\lambda}\}$, and the trace term $\mathrm{tr}(P_{\mathbf{1}}^{\perp}\tilde{R}H^{-1}\tilde{R}^TP_{\mathbf{1}}^{\perp})$. Plotted in the left and middle frames of Figure 1 are the $\mathrm{KL}(\eta, \eta_\lambda)$ of the cross-validated λ's with $\alpha = 1$ and $\alpha = 1.4$ versus the minimum $\mathrm{KL}(\eta, \eta_\lambda)$ on the grid. The relative efficacy of the methods, defined as the ratio of the horizontal axis to the vertical axis, are shown in the right frame in box plots along with those of $\alpha = 1.2, 1.6$. Figure 1 is virtually a duplicate of Figure 6.1 in Gu (2002), included here for a convenient comparison with the bivariate and trivariate simulation results reported below; further discussions concerning the univariate simulation can be found in Gu (2002, §6.3.3), where, among other things, it was shown that the performance of indirect cross-validation is comparable to that of (5) with $\alpha = 1$.

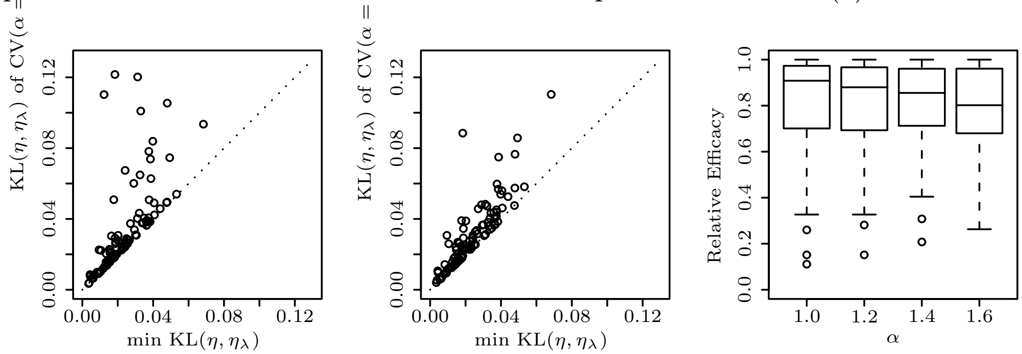

Figure 1. Performance of cross-validation in univariate simulation. Left: Loss achieved by cross-validation with $\alpha = 1$. Center: Loss achieved by cross-validation with $\alpha = 1.4$. Right: Relative efficacy of cross-validation with $\alpha = 1, 1.2, 1.4, 1.6$.

Now consider a bivariate test density

$$f_2(x, y) \propto f_1(x)e^{-12.5(y-0.5)^2}I_{[0<x<1,0<y<1,x+y<1]}, \tag{7}$$

which is supported on a triangular domain $\mathcal{X} = \{(x, y) : x, y > 0, x + y < 1\}$; $f_1(x)$ is as given in (6). Samples of size $n = 300$ were drawn from $f_2(x, y)$, and the additive model $\eta(x,y)=\eta_x+\eta_y$ was fitted to the data with $J(\eta)=\theta_x^{-1}\int_0^1\ddot{\eta}_x^2dx+\theta_y^{-1}\int_0^1\ddot{\eta}_y^2dy$; $\mathcal{N}_J=\{(x-0.5),(y-0.5)\}$ and $R_J((x,y),(u,v))=\theta_x\{k_2(x)k_2(u)-k_4(|x-u|)\}+\theta_y\{k_2(y)k_2(v)-k_4(|y-v|)\}$. With the two smoothing parameters of the additive model, it is still feasible to lay a grid and the triangular domain keeps things nontrivial. See §6 for a truncated domain in real-data application. Estimates were calculated with $q=100$ and $(\log_{10}(\lambda/\theta_x),\log_{10}(\lambda/\theta_y))$ over a 21×21 regular grid on $[-7,-3]\times[-6,-2]$, for 100 replicates; the choice of q is discussed in §4.2. Recorded for each of the estimates were the Kullback-Leibler distance $\mathrm{KL}(\eta,\eta_\lambda)$, the minus log likelihood $n^{-1}\sum_{i=1}^n\{-\eta_\lambda(X_i)+\log\int_{\mathcal{X}}e^{\eta_\lambda}\}$, and the trace term $\mathrm{tr}(P_{\mathbf{1}}^\perp\tilde{R}H^{-1}\tilde{R}^TP_{\mathbf{1}}^\perp)$. Parallel to Figure 1, the bivariate simulation results are summarized in Figure 2.

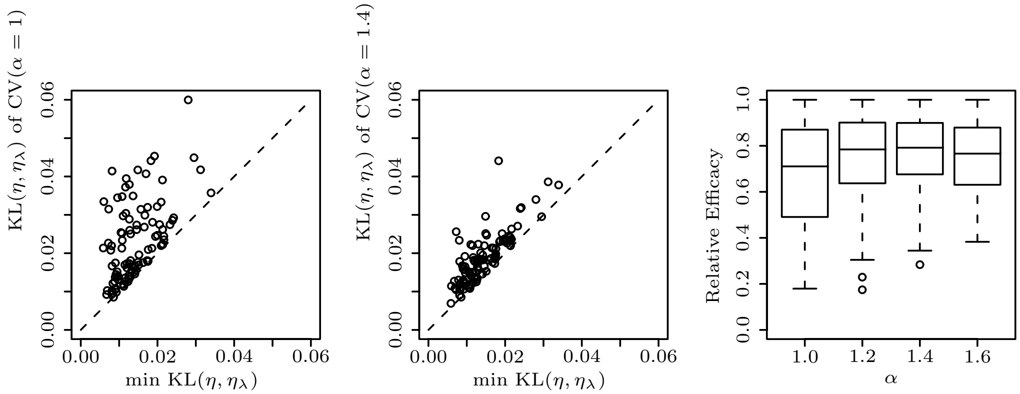

Figure 2. Performance of cross-validation and modifications thereof in bivariate simulation. Left: Loss achieved by cross-validation with $\alpha=1$. Center: Loss achieved by cross-validation with $\alpha=1.4$. Right: Relative efficacy of cross-validation with $\alpha=1,1.2,1.4,1.6$.

Finally consider a trivariate test density

$$f_3(x,y)\propto f_1(x-0.3z+0.1)f_1(y-0.2z+0.1)e^{-12.5(z-0.5)^2}I_{[0<x,y,z<1]},\qquad(8)$$

where $f_1(x)$ is as given in (6). Data were generated from $f_3(x,y,z)$, and a model $\eta(x,y,z)=\eta_x+\eta_y+\eta_z+\eta_{x,z}+\eta_{y,z}$ was fitted to the data; note the conditional independence of X and Y given Z built into the model. Tensor product cubic splines were calculated with $\mathcal{N}_J=\{k_1(x),k_1(y),k_1(z),k_1(x)k_1(z),k_1(y)k_1(z)\}$ and

$$\begin{aligned}&R_J((x,y,z),(u,v,w))\\&=\theta_xR_c(x,u)+\theta_yR_c(y,v)+\theta_zR_c(z,w)\\&\quad+\theta_{x,z}^{(1)}R_c(x,u)k_1(z)k_1(w)+\theta_{x,z}^{(2)}k_1(x)k_1(u)R_c(z,w)+\theta_{x,z}^{(3)}R_c(x,u)R_c(z,w)\\&\quad+\theta_{y,z}^{(1)}R_c(y,v)k_1(z)k_1(w)+\theta_{y,z}^{(2)}k_1(y)k_1(v)R_c(z,w)+\theta_{y,z}^{(3)}R_c(y,v)R_c(z,w),\end{aligned}$$

where $R_c(x,u) = k_2(x)k_2(u) - k_4(|x-u|)$ and the rest of the notation can be found in §2; the expression of the penalty $J(\eta)$, which also has 9 terms, is omitted here. One hundred replicates of samples of size $n = 300$ were drawn, and estimates were calculated with $q = 100$ and with the smoothing parameters minimizing the loss $\mathrm{KL}(\eta, \eta_\lambda)$ and the cross-validation score with $\alpha = 1, 1.2, 1.4, 1.6$; separate minimizations were conducted as a grid search was no longer feasible. Recorded for each of the estimates were the Kullback-Leibler distance $\mathrm{KL}(\eta, \eta_\lambda)$. Parallel to Figures 1 and 2, the simulation results are summarized in Figure 3.

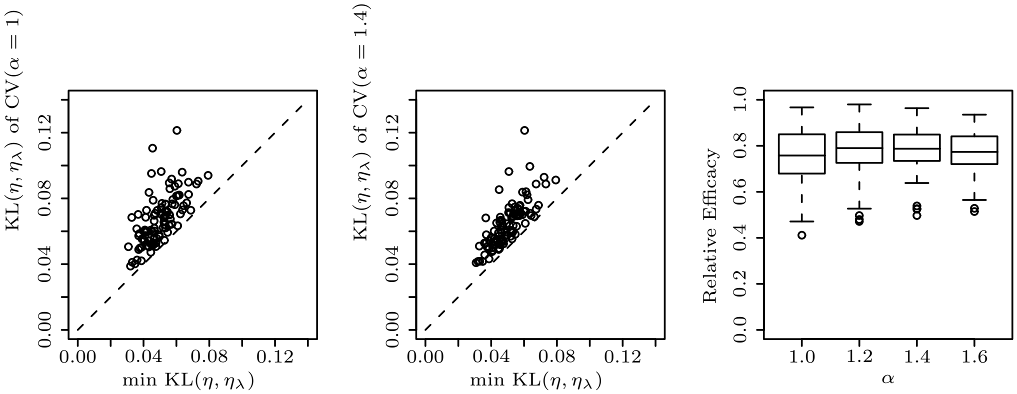

Figure 3. Performance of cross-validation and modifications thereof in trivariate simulation. Left: Loss achieved by cross-validation with $\alpha = 1$. Center: Loss achieved by cross-validation with $\alpha = 1.4$. Right: Relative efficacy of cross-validation with $\alpha = 1, 1.2, 1.4, 1.6$.

4.2. Empirical choice of q

As mentioned in §2, a dimension of the order $q \asymp n^{2/(pr+1)+\epsilon}$, $\forall \epsilon > 0$, is sufficient for asymptotic efficiency. It can be shown that $r = 4$ for the univariate cubic spline and the cubic spline additive model, and $r = 4 + \delta$, $\forall \delta > 0$, for tensor product cubic splines with interactions; see, e.g., Gu (1996). Since $\epsilon, \delta > 0$ can be arbitrarily small, one may use $q = kn^{2/(4p+1)}$ in practice. We present some simulation results to suggest adequate values of k for practical use.

Consider the test densities $f_1(x)$ of (6) and $f_2(x,y)$ of (7); the test densities are sufficiently smooth so $p = 2$. Samples of sizes $n = 150, 300, 600$ were drawn from $f_1(x)$ and $f_2(x,y)$, respectively. For each of the six samples and every k on the grid 1(1)15, 30 different random subsets $\{Z_j\} \subset \{X_i\}$ of size $q = kn^{2/9}$ were generated to form 30 different $\mathcal{H}_q$, and 30 different estimates were calculated based on the same data with the smoothing parameters selected by cross-validation with $\alpha = 1.4$. The Kullback-Leibler losses of the 30 estimates were calculated and summarized in boxplots. Shown in the left and center frames of Figure 4 are the boxplots for $k = 5(1)15$; the first four are much wider and their inclusion would greatly reduce the resolution of the ones shown. The fact that the box width gradually decreases as k increases indicates that $q \asymp n^{2/9}$ indeed appears to be the "correct" scale. The plots suggest that a k as small as 8 or 9 could be stable enough for practical use. Similar plots on the $q \asymp n^{2/5}$ scale (not shown here) have also been inspected, and the much faster shrinking rate of the boxes suggests that $q \asymp n^{2/5}$ may not be the proper scale.

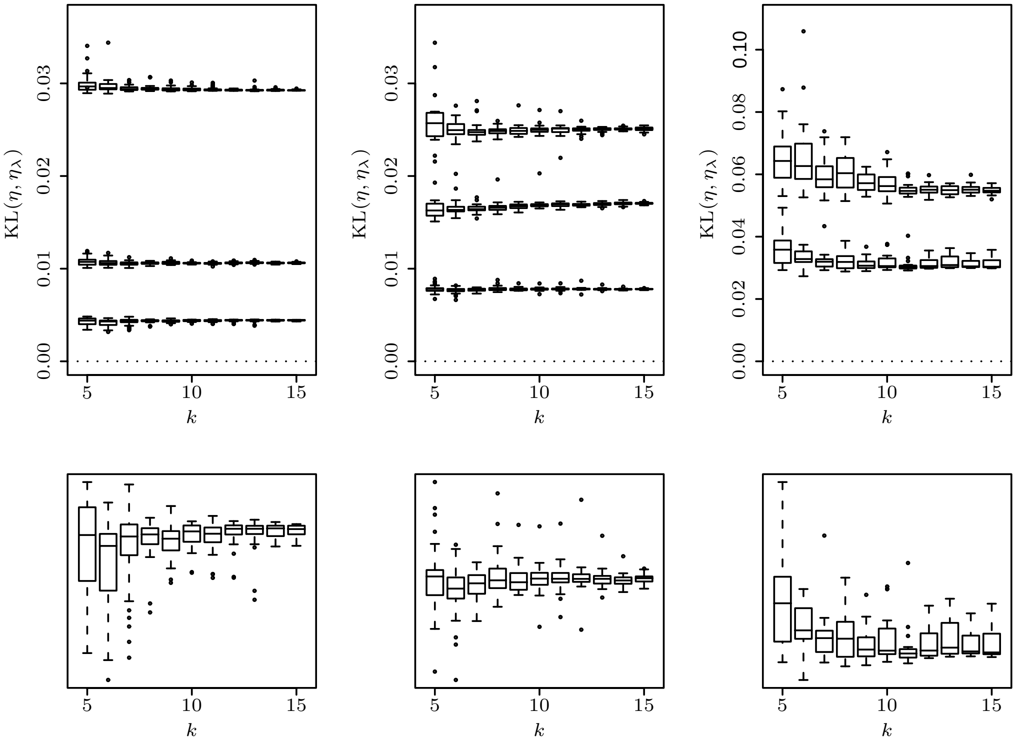

Figure 4. Effect of q on estimation consistency. Left: $f_1(x)$. Center: $f_2(x, y)$. Right: $f_3(x, y, z)$. Boxplots of 30 (Left and Center) or 20 (Right) $\mathrm{KL}(\eta, \eta_\lambda)$ for each of $q = kn^{2/9}$. Top: from high to low, $n = 150, 300, 600$ (Left and Center) or $n = 300, 600$ (Right). Bottom: $n = 600$.

Samples of sizes $n = 300, 600$ were also drawn from $f_3(x, y, z)$ of (8), and the above experiment was repeated but with only 20 different subsets $\{Z_j\} \subset \{X_i\}$ for each of $q = kn^{2/9}$; the results are summarized in the right frame of Figure 4. The numerical inconsistency appears to be much higher than the level seen in the univariate and bivariate simulations. Similar plots on the $q = kn^{2/7}$ scale were also inspected (not shown here), where the $n = 300$ fits gradually settled down, as hinted by the trend seen in the right frame of Figure 4, but the inconsistency of the $n = 600$ fits remained at the "constant" level seen. A major contributor to the high level of numerical inconsistency is likely the large number of smoothing parameters (nine total). The consistency in terms of the general performance appears tolerable.

In practice, we suggest the use of $q = kn^{2/9}$ with k around 10 for tensor product cubic splines. We were not able to find an example with "barely" square integrable second derivatives, and we doubt there are many such "true" functions in the real world. Since the computation is so much faster (some timing results can be found in §6), quick checks on the stability can be performed simply by comparing estimates with different subsets $\{Z_j\} \subset \{X_i\}$.

For the bivariate and trivariate simulations of §4.1 with $n = 300$, we used $q = 100 \approx 10n^{2/5} \approx 28n^{2/9}$. Our purpose there was to study the behavior of cross-validation, and we chose a q sufficiently large to ensure stability, yet small enough so the experiments were feasible within reasonable time.

In a separate study, Wahba, Lin and Leng (2002) reported a stable value of $q = 40$ for $n = 1000$; note that $40 \approx 8.6(1000)^{2/9}$.

5. Numerical Integration

Numerical integration is an essential part of the method under study. We briefly discuss a few practical issues concerning the use of quadratures and cubatures in the setting.

For the calculation of $\int_{\mathcal{X}} g(x)d\nu(x)$, a quadrature or cubature is of the form $\sum_{i=1}^{d} w_i g(x_i)$, where x_i are the nodes and w_i are the associated weights; typically, one dimensional formulas are called quadratures and multidimensional ones are called cubatures. Within a family of formulas, the accuracy usually increases with the number of nodes, or size, along with the cost. Certain methods are adaptive, attempting to achieve user-specified precision through sequential node addition guided by precision estimates.

In our setting, $O(q^2)$ integrals involving the same $O(q)$ functions need to be calculated for each step of the Newton iteration, so formulas with fixed nodes are actually more economical than the adaptive methods. Also, high accuracy is not as essential in the fitting stage as such consistencies as the non-negative definiteness of H; H is guaranteed to be non-negative definite with fixed nodes and positive weights.

In one dimension, a standard Gauss quadrature with d up to 200 is sufficient for our needs. The public domain subroutine `gaussq.f` from `netlib.org` can be used to generate the nodes and the weights.

For the bivariate simulation, we used a naive flat-weight 40 × 40 regular grid truncated to the triangular domain, with the diagonal nodes carrying half weights. We were not able to find a more efficient formula on the triangular domain.

On multidimensional cubes, product quadratures quickly become prohibitive. A system known as Smolyak algorithm has been developed in the literature for the derivation of efficient cubatures from univariate formulas. The efficiency of Smolyak cubatures is achieved by thinning out nodes from the product quadratures; some negative weights are introduced in the process. Some of the Smolyak cubatures can be found in Novak and Ritter (1996) and Petras (2001). Public domain routines from Knut Petras' SMOLPACK can be modified to return the nodes and the weights of Smolyak cubatures.

Smolyak cubatures are highly accurate for smooth functions, but we have experienced great difficulty applying them in our problem without modification. The data are typically away from the boundaries of the domain one specifies, but the placement of nodes in Smolyak cubatures is dense near the boundaries and sparse in the middle, causing them to miss the peaks in the intermediate and final estimates $e^{\tilde{\eta}}/\int_{\mathcal{X}} e^{\tilde{\eta}}$, resulting in gross under-approximations of the true integrals. To circumvent the problem, we apply transformations on each coordinate of the cube to make the marginal data nearly uniformly distributed, then use the Smolyak formulas on the transformed domain.

To illustrate the strategy, consider integration on $[0,1]^2$. First estimate the marginal densities $f_x(x)$ and $f_y(y)$ with distribution functions F_x and F_y; a bit oversmoothing does no harm for the purpose so we use cross-validation with $\alpha = 2$. Transforming the domain by $\tilde{x} = F_x(x)$ and $\tilde{y} = F_y(y)$, the marginal observations are nearly uniformly distributed on the $\tilde{x}$ and $\tilde{y}$ scales. Let $(\tilde{x}_i, \tilde{y}_i)$ be the Smolyak nodes and w_i be the corresponding weights, the integral $\int_0^1\int_0^1 g(x,y)dxdy = \int_0^1\int_0^1 g(F_x^{-1}(\tilde{x}), F_y^{-1}(\tilde{y}))(dx/d\tilde{x})(dy/d\tilde{y})d\tilde{x}d\tilde{y}$ can be approximated by

$$\sum_{i=1}^{d} \frac{w_i g(F_x^{-1}(\tilde{x}_i), F_y^{-1}(\tilde{y}_i))}{f_x(F_x^{-1}(\tilde{x}_i)) f_y(F_y^{-1}(\tilde{y}_i))},$$

where $f_x(F_x^{-1}(\tilde{x})) = d\tilde{x}/dx$ and $f_y(F_y^{-1}(\tilde{y})) = d\tilde{y}/dy$. An example of this is shown in Figure 5, where the circles are 150 simulated observations and the filled dots are the nodes of the 449-point version of the so-called delayed Smolyak cubature in two dimension, on the original scale and on the transformed scale; the transformations are through the marginal density estimates based on the 150 observations.

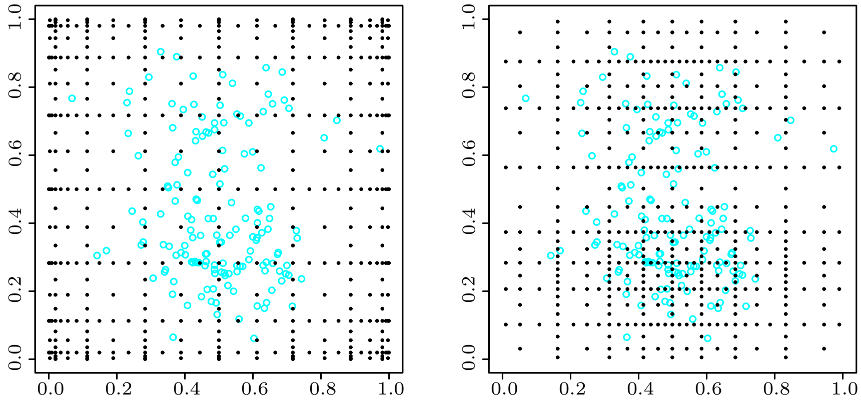

Figure 5. Smolyak cubature in two dimension. Left: Original scale. Right: Transformed scale. Circles are the data and filled dots are cubature nodes.

6. Example

We now apply the techniques developed above to analyze a data set. To study the AIDS incubation time, a valuable source of information is the records of patients who were infected with the HIV virus through blood transfusion since the date can be ascertained retrospectively. A data set collected by the Centers for Disease Control and Prevention (CDC) is listed in Wang (1989), which includes the time X from the transfusion to the diagnosis of AIDS, the time Y from the transfusion to the end of study (July 1986), both in months, and the age of the individual at the time of transfusion, for 295 individuals. It is clear that $X \leq Y$.

Assuming the independence of X and Y in the absence of truncation, and conditioning on the truncation mechanism, an additive model of the log density can be fitted using the formulation in the bivariate simulations of §4 but with the domain replaced by $\mathcal{X} = \{(x,y) : 0 \leq x \leq y \leq 100\}$. Using cross-validation with $\alpha = 1.4$ and $q = n = 295$, the estimated density $f(x,y) = e^{\eta_x+\eta_y}/\int_{\mathcal{X}} e^{\eta_x+\eta_y}$ is contoured in Figure 6 as solid lines, with the marginal densities $f(x) = e^{\eta_x}/\int_0^{100} e^{\eta_x}$ and $f(y) = e^{\eta_y}/\int_0^{100} e^{\eta_y}$ plotted in the empty space and the data superimposed. With $q = 28 \approx 8n^{2/9}$, a fit is superimposed in Figure 6 as dashed lines. The estimated marginal densities differ slightly on the upper end of $f(x)$ and the lower end of $f(y)$, where data are scarce due to truncation. Further analyses of the data set can be found in Gu (2002, §1.4.2, §6.5.3, §6.6.4).

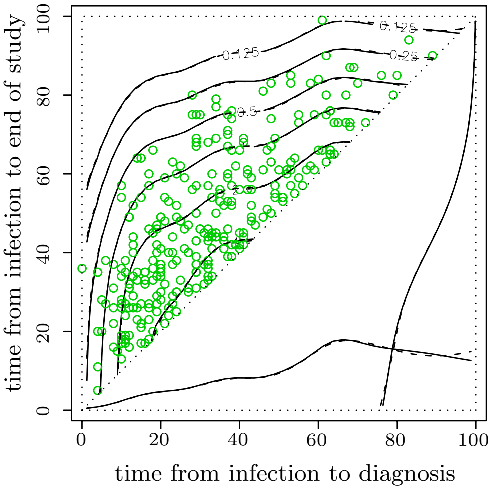

Figure 6. AIDS incubation and HIV infection. Contours are estimated density on the observable region surrounded by dotted lines. Circles are the observations. Curves over the dotted lines in the empty space are the estimated marginal densities. The solid lines are for $q = 295$ and the dashed lines are for $q = 28$; they are nearly indistinguishable.

On a workstation with dual Athlon MP1800+ and 2GB RAMS running FreeBSD 4.4 and R 1.4.0, the computation for $q = 295$ took about 149 CPU seconds and that for $q = 28$ took less than 3 CPU seconds.

7. Remarks

In this article, we have developed an effective and efficient implementation of the penalized likelihood method for probability density estimation. The techniques have been coded into a set of R functions, which were used to calculate the numerical examples. A polished user-interface can be found in the `ssden` suite in the R package `gss` by the first author.

The empirical rules of $\alpha \approx 1.4$ for cross-validation and $q \approx 10n^{2/9}$ for scalable approximation are practically convenient. Many more examples had been looked at besides those presented, and it was remarkable that some systematic pattern emerged. Kim and Gu (2002) also observed nearly identical rules in the Gaussian regression setting.

The direct cross-validation strategy also applies to penalized hazard estimation with little modification. Some empirical results can be found in Gu (2002, Chap. 7).

The current development settles the practical computability of the method developed in Gu and Qiu (1993), at least for dimensions up to 3 or 4 where the cubature sizes are manageable. Equipped with the computational tool, work is under way for the development of tools for the assessment of the significance of model terms in log density, which often have (conditional) independence implications. With little modification, the tools can also be applied in closely related settings such as the conditional density estimation of Gu (1995).

Appendix. Convergence Rates of Estimates

In this appendix, we sketch the asymptotic theory developed in Gu and Qiu (1993) and refined in Gu (2002, §8.2), and point out the critical link that leads to the justification for the use of $\mathcal{H}_q$ mentioned in §2.

Let $e^\eta / \int_\mathcal{X} e^\eta$ be the density from which the data are generated. Define $V(g) = \mu_\eta(g^2) - \mu_\eta^2(g)$, where $\mu_\eta(g) = \int_\mathcal{X} g e^\eta / \int_\mathcal{X} e^\eta$. The asymptotic convergence rates of the minimizer $\hat{\eta}$ of (1) in $\mathcal{H}$ are governed by an eigenvalue analysis of $V(g)$ with respect to $J(g)$. Under mild conditions, it can be shown that there exist eigenfunctions ϕ_ν such that $V(\phi_\nu, \phi_\mu) = \delta_{\nu,\mu}$, $J(\phi_\nu, \phi_\mu) = \rho_\nu \delta_{\nu,\mu}$, where $\delta_{\nu,\mu}$ is the Kronecker delta, and $\rho_\nu > C\nu^r$ for some $r > 1$, $C > 0$, and ν sufficiently large; $V(g,h)$ and $J(g,h)$ are the inner products associated with $V(g)$ and $J(g)$, respectively. It then can be shown that as $\lambda \to 0$ and $n\lambda^{1/r} \to \infty$, $V(\hat{\eta} - \eta) = O_p(\lambda^p + n^{-1}\lambda^{-1/r})$ and $\mathrm{SKL}(\eta, \hat{\eta}) = O_p(\lambda^p + n^{-1}\lambda^{-1/r})$ for some $p \in [1,2]$, where $\mathrm{SKL}(\eta, \hat{\eta}) = \mathrm{KL}(\eta, \hat{\eta}) + \mathrm{KL}(\hat{\eta}, \eta)$ is the symmetrized Kullback-Leibler between η and $\hat{\eta}$. The constant p depends on how smooth η is: the rates given hold for η satisfying $\sum_\nu \rho_\nu^p \eta_\nu^2 < \infty$ for p up to 2, where $\eta_\nu = V(\eta, \phi_\nu)$; note that $\sum_\nu \rho_\nu \eta_\nu^2 = J(\eta)$. The optimal rates are given by $O_p(n^{-pr/(pr+1)})$, achieved with $\lambda \asymp n^{-r/(pr+1)}$.

Let $\mathcal{H}^*$ be a subspace of $\mathcal{H}$ such that $V(h) = O_p(\lambda J(h))$, $\forall h \in \mathcal{H} \ominus \mathcal{H}^*$. By the analysis of Gu and Qiu (1993, §6), it can be shown that the same convergence rates hold for the minimizer of (1) in $\mathcal{H}^*$. By the proof of Lemma 6.1 in Gu and Qiu (1993) one has, for $h \in \mathcal{H} \ominus \mathcal{H}_q$,

$$V(h) = O_p(q^{-1/2}\lambda^{-1/r})(V + \lambda J)(h). \tag{9}$$

Taking $q \asymp n^{2/(pr+1)+\epsilon}$ and $\lambda \asymp n^{-r/(pr+1)}$, $\forall \epsilon > 0$, one has $V(h) = o_p(\lambda J(h))$.

The stated result of Lemma 6.1 in Gu and Qiu (1993) concerns only the case with $q = n$, but the proof actually establishes the result quoted in (9).

Acknowledgements

This research was supported by National Institutes of Health under Grant R33HL68515, and by Purdue Research Foundation under a research grant.

References

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

Craven, P. and G. Wahba (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.

Dennis, J. E. and R. B. Schnabel (1996). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations.* SIAM, Philadelphia. Corrected reprint of the 1983 original.

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

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

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

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

Gu, C. (1998). Structural multivariate function estimation: Some automatic density and hazard estimates. *Statist. Sinica* **8**, 317-335.

Gu, C. (2002). *Smoothing Spline ANOVA Models.* Springer-Verlag, New York.

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

Kim, Y.-J. and C. Gu (2002). Penalized least squares regression: fast computation via efficient approximation. Technical report, Department of Statistics, Purdue University, West Lafayette, IN.

Leonard, T. (1978). Density estimation, stochastic processes and prior information. *J. Roy. Statist. Soc. Ser. B* **73**, 113-146 (with discussions).

Novak, E. and K. Ritter (1996). High dimensional integration of smooth functions over cubes. *Numer. Math.* **75**, 79-97.

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

Petras, K. (2001). Asymptotically minimal smolyak cubature. Preprint.

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

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

Wahba, G., Y. Lin and C. Leng (2002). Comment on "Spline adaptation in extended linear models" by M. H. Hansen and C. Kooperberg. *Statist. Sci.* **17**, 33-37.

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

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

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

E-mail: chong@stat.purdue.edu

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

E-mail: jingyuan@stat.purdue.edu

(Received May 2002; accepted April 2003)