SMOOTHING SPLINE DENSITY ESTIMATION UNDER BIASED SAMPLING

by

Chong Gu
Purdue University

Technical Report #92-03

Department of Statistics
Purdue University

January 1992

◇

Smoothing Spline Density Estimation Under Biased Sampling

CHONG GU*

Purdue University

Abstract

This article extends the smoothing spline probability density estimation techniques developed by Gu and Qiu (1991) and Gu (1991) to Vardi's (1985) selection bias models. The estimation is via estimating the log density in a reproducing kernel Hilbert space by the standard penalized likelihood method. The existence of the estimator is discussed and the asymptotic convergence rates in an properly defined symmetrized Kullback-Leibler and in a related weighted mean square error are obtained. A computable adaptive semiparametric estimator is proposed which shares with the original estimator the same asymptotic convergence rates. The computation of the estimator with an automatic smoothing parameter is also discussed. A simulation study is presented to illustrate the relative effectiveness of the automatic smoothing parameter selection and the best and the worst cases in the simulations are presented to illustrate the absolute performance of the techniques.

AMS 1991 subject classifications. Primary 62G07; secondary 65D07, 65D10, 41A25, 41A65.

Key words and phrases. Biased sampling, density estimation, symmetrized Kullback-Leibler, penalized likelihood, rate of convergence, semiparametric estimator, smoothing parameter.

1 Introduction

Let X_i, $i = 1, \cdots, n$, be independent observations on a domain $\mathcal{X}$ sampled from probability densities proportional to $w_i(x)f(x)$, where $w_i \geq 0$ are known biasing functions and f is an unknown probability density assumed to be "smooth". The purpose of this article is to propose and study a

*Research supported by NSF Grant DMS-9101730.

penalized likelihood estimator of f based on the data (w_i, X_i). Let $\mathcal{T}$ be an index set and $w(t,x)$ a known function on $\mathcal{T} \times \mathcal{X}$ such that the set $\{w(t,\cdot), t \in \mathcal{T}\}$ includes all possible biasing functions and $w(t,\cdot) \neq w(t',\cdot)$ when $t \neq t'$. The "observed" biasing function w_i can then be denoted as $w(t_i,\cdot)$ for some $t_i \in \mathcal{T}$ and the data are now (t_i, X_i). Assume $0 < \int_{\mathcal{X}} w(t,x)f(x)dx < \infty$, $\forall t \in \mathcal{T}$, so that the densities $w(t,x)f(x)/\int_{\mathcal{X}} w(t,x)f(x)$ are well defined. Take t_i as observations from a probability density $m(t)$ on $\mathcal{T}$. The data (t_i, X_i) can then be treated as from a two-stage sampling.

Example 1.1 *Ordinary sampling.* Let $\mathcal{T} = \{1\}$ and $w(1,x) = 1$. X_i are *i.i.d.* observations from $f(x)$. □

Example 1.2 *Length-biased sampling.* Let $\mathcal{T} = \{1\}$, $\mathcal{X} = [0,1]$, and $w(1,x) = x$. X_i are *i.i.d.* length-biased observations from the probability density $xf(x)/\int_0^1 xf(x)$. □

Example 1.3 *Ordinary and length-biased sampling.* Let $\mathcal{T} = \{1,2\}$, $\mathcal{X} = [0,1]$, $w(1,x) = 1$, and $w(2,x) = x$. $X_i|t_i = 1$ are ordinary observations from $f(x)$ and $X_i|t_i = 2$ are length-biased observations from $xf(x)/\int_0^1 xf(x)$. Examples 1.1 and 1.2 are special cases with $m(1) = 1$ and $m(1) = 0$, respectively. □

Example 1.4 *Finite "strata" biased sampling.* Let $\mathcal{T} = \{1, \cdots, s\}$ and $\mathcal{X} = \overline{\cup_{t:m(t)>0}\{x : w(t,x) > 0\}}$, where $w(t,x) \geq 0$ but otherwise arbitrary. $X_i|t_i$ are from $w(t_i,x)f(x)/\int_{\mathcal{X}} w(t_i,x)f(x)$. Example 1.3 is a special case with $s = 2$. □

An early reference on length-biased sampling and its applications is Cox (1969). The empirical distribution for Example 1.3 was derived and its asymptotic properties were studied by Vardi (1982). Vardi (1985) further investigated the existence and uniqueness of the empirical distribution in the more general models of Example 1.4 and derived an algorithm to compute it when it exists. Gill, Vardi, and Wellner (1988) developed a large sample theory for Vardi's (1985) empirical distribution. Practical applications of the biased sampling models are discussed by the above authors and further references cited by them. Note that the probability density corresponding to the empirical distribution is a delta sum and hence is ultimately rough. Some attempts have been made to smooth the empirical distribution via the kernel method when $\mathcal{T}$ is a singleton; see, e.g., Jones (1991).

For the ordinary sampling model of Example 1.1, Good and Gaskins (1971) proposed to estimate f via the penalized likelihood method. Further developments on the line have been made by Leonard (1978) and Silverman (1982), among others, and recently by Gu and Qiu (1991) and Gu (1991). In this article, I shall apply the techniques emerged in the recent developments to study the estimation of f in the biased sampling models.

The penalized likelihood method estimates f by the minimizer of a functional $L(f|\text{data})+\lambda J(f)$, where $L(f|\text{data})$ measures the lack of fit, $J(f)$ measures the roughness, and $\lambda > 0$ controls the tradeoff. Assume $f > 0$ on $\mathcal{X}$. Leonard (1978) introduced the logistic density transform $f = e^g/\int_{\mathcal{X}} e^g$ where g is to be estimated, which is constraint-free and assures the estimate to be a genuine density. To make the transform one-to-one, Gu and Qiu (1991) excluded the space span$\{1\}$ of constant functions from the model space of g. The smoothing spline estimate of f for ordinary sampling defined by Gu and Qiu (1991) is thus $e^{\hat{g}}/\int_{\mathcal{X}} e^{\hat{g}}$ where the $\hat{g}$ is the minimizer of

$$-\frac{1}{n}\sum_{i=1}^{n} g(X_i) + \log\int_{\mathcal{X}} e^g + \frac{\lambda}{2}J(g) \tag{1.1}$$

subject to $g \in \mathcal{H}$, where $\mathcal{H} \not\supset \text{span}\{1\}$ is a Hilbert space of functions on $\mathcal{X}$ and J is a square (semi)norm in $\mathcal{H}$ with a finite dimensional null space $J_\perp \subset \mathcal{H}$. The first term in (1.1) is simply the standard minus log likelihood. For (1.1) to be well defined at $g = 0$, $\mathcal{X}$ has to be finite. For the first term of (1.1) to be continuous in g, it shall be assumed that an evaluation $[x]g = g(x)$ is continuous in $\mathcal{H}$. A Hilbert space in which evaluation is continuous is known to be a reproducing kernel Hilbert space possessing a reproducing kernel $R(x, y)$, a positive-definite bivariate function on $\mathcal{X} \times \mathcal{X}$, such that $\text{span}\{R(x,\cdot), x \in \mathcal{X}\} = \mathcal{H}$ and $\langle R(x,\cdot), g(\cdot)\rangle = f(x)$; see Aronszajn (1950). Since $\langle R(x,\cdot), R(x,\cdot)\rangle = R(x,x)$, $|[x]g| \le R^{1/2}(x,x)\|g\|$ by Cauchy-Schwartz and the equality holds when $g = R(x,\cdot)$, so the norm of $[x]$ is $R^{1/2}(x,x)$. When $R(x,y)$ is continuous, $g \in \mathcal{H}$ is continuous, e^g Riemann integrable, and $R(x,x)$ bounded on a bounded $\mathcal{X}$, and in turn the second term of (1.1) is continuous in g via the Riemann sum approximation of the integrals. By estimating g in $\mathcal{H}$, one implicitly assumes that the truth g_0 is a member of $\mathcal{H}$, whose smoothness is characterized by the square (semi)norm J. Let $\text{SKL}(g,h)$ be the symmetrized Kullback-Leibler between $e^g/\int_{\mathcal{X}} e^g$ and $e^h/\int_{\mathcal{X}} e^h$. Under appropriate conditions, Gu and Qiu (1991) obtained the asymptotic convergence rate of $\hat{g}$ in $\text{SKL}(\hat{g}, g_0)$ and derived an adaptive semiparametric estimator $\hat{g}_n$ such that $\text{SKL}(\hat{g}_n, g_0)$ shares the same rate as $\text{SKL}(\hat{g}, g_0)$. Gu (1991) developed an automatic algorithm to calculate $\hat{g}_n$

with a properly chosen λ.

Note that although the integration measure in $\int_{\mathcal{X}} e^g$ of (1.1) is usually understood as the uniform measure on $\mathcal{X}$, neither the theory of Gu and Qiu (1991) nor the algorithm of Gu (1991) discriminate against other measures. By the chain rule of the Radon-Nikodym derivative, the biased samples from $w(x)f(x)$ under the uniform integration measure are simply ordinary samples from $f(x)$ under the integration measure $\nu_w(A) = \int_A w(x)dx$. So for a singleton $\mathcal{T}$ such as the length-biased model of Example 1.2 there is nothing to be done. The remaining of this article documents the extensions of the theory and the algorithm to a general $\mathcal{T}$ where one has to combine information from different types of sources. The development parallels those of Gu and Qiu (1991) and Gu (1991) and is organized as follows. In Section 2, I shall define the penalized likelihood estimator $\hat{g}$ and discuss its existence, define a symmetrized Kullback-Leibler $\mathrm{SKL}(g, h)$ properly modified according to the sampling structure, and discuss the smoothness assumptions. In Section 3, $\mathrm{SKL}(\hat{g}, g_0)$ is calculated. In Section 4, a semiparametric $\hat{g}_n$ is proposed and $\mathrm{SKL}(\hat{g}_n, g_0)$ calculated. Section 5 discusses the calculation of $\hat{g}_n$ with an automatic λ and Section 6 presents simulation results.

2 Penalized Likelihood Estimation

Assume $f > 0$ on a bounded $\mathcal{X}$. Based on independent observations (t_i, X_i), $i = 1, \cdots, n$, where $X_i|t_i \sim w(t_i, x)f(x)/\int_{\mathcal{X}} w(t_i, x)f(x)$, the likelihood of $f = e^g/\int_{\mathcal{X}} e^g$ is

$$\prod_{i=1}^{n} \{w(t_i, X_i)e^{g(X_i)}/\int_{\mathcal{X}} w(t_i, x)e^{g(x)}\}. \tag{2.1}$$

Define the penalized likelihood estimator of f as $e^{\hat{g}}/\int_{\mathcal{X}} e^{\hat{g}}$, where the $\hat{g}$ minimizes

$$P_\lambda(g) = -\frac{1}{n}\sum_{i=1}^{n} g(X_i) + \frac{1}{n}\sum_{i=1}^{n} \log \int_{\mathcal{X}} w(t_i, x)e^{g(x)} + \frac{\lambda}{2}J(g), \tag{2.2}$$

subject to $g \in \mathcal{H} \not\supset \mathrm{span}\{1\}$, and $\mathcal{H}$ is a reproducing kernel Hilbert space with a continuous reproducing kernel R and J is a (semi)norm in $\mathcal{H}$ with a finite dimensional null space $J_\perp$. The functional (2.2) reduces to (1.1) for the ordinary sampling model of Example 1.1. Examples of $(\mathcal{X}, \mathcal{H}, J)$ can be found in Gu and Qiu (1991, §2).

Theorem 3.1 of Gu and Qiu (1991) states that, if the likelihood part of (2.2), $L(g)$ say, is continuous and strictly convex in $g \in \mathcal{H}$, then the minimizer of (2.2) in $\mathcal{H}$ exists whenever it exists

in $J_\perp$. For $L(g)$ to be continuous in g, it suffices to further assume that $w(t_i, x)$ is bounded. For $L(g)$ to be strictly convex in $\mathcal{H} \not\supset \text{span}\{1\}$, it is necessary and sufficient (by Holder's inequality) to have $\mathcal{X} = \overline{\cup_{1\le i\le n}\{x : w(t_i, x) > 0\}}$, which essentially means that the data do carry information about f on the whole domain $\mathcal{X}$.

I now derive an appropriate score for assessing the estimation precision. It is easy to verify that the symmetrized Kullback-Leibler between $we^g/\int_{\mathcal{X}} we^g$ and $we^h/\int_{\mathcal{X}} we^h$ is

$$\text{SKL}_w(g, h) = \int_{\mathcal{X}} (g - h)\left(\frac{we^g}{\int_{\mathcal{X}} we^g} - \frac{we^h}{\int_{\mathcal{X}} we^h}\right).$$

Given data from $we^{g_o}/\int_{\mathcal{X}} we^{g_o}$, $\text{SKL}_w(\hat{g}, g_0)$ defines a proper measure for the estimation precision. In the general two stage sampling setup,

$$\text{SKL}(g, h) = \int_{\mathcal{T}} m(t) \int_{\mathcal{X}} (g(x) - h(x))\left(\frac{w(t, x)e^{g(x)}}{\int_{\mathcal{X}} w(t, x)e^{g(x)}} - \frac{w(t, x)e^{h(x)}}{\int_{\mathcal{X}} w(t, x)e^{h(x)}}\right) \tag{2.3}$$

defines a weighted average of SKL_w's with the weight function $m(t)$ proportional to the resources allocated to $w(t, x)e^{g_o(x)}/\int_{\mathcal{X}} w(t, x)e^{g_o(x)}$, and hence $\text{SKL}(\hat{g}, g_0)$ makes an adequate criterion for assessing the quality of $e^{\hat{g}}/\int_{\mathcal{X}} e^{\hat{g}}$ as an estimator of $e^{g_o}/\int_{\mathcal{X}} e^{g_o}$ under the biased sampling structure.

Let $\mu_g(h) = \int_{\mathcal{T}} m(t)\mu_g(h|t)$ where $\mu_g(h|t) = \int_{\mathcal{X}} h(x)w(t, x)e^{g(x)}/\int_{\mathcal{X}} w(t, x)e^{g(x)}$. Using the boundedness of $R(x, x)$ and the Riemann sum approximation of integrals, it can be shown that $\mu_{g+\alpha f}(h)$ is continuously differentiable as a function of the scalar α for $g, f, h \in \mathcal{H}$, and

$$\frac{d\mu_{g+\alpha f}(h)}{d\alpha} = \int_{\mathcal{T}} m(t)\{\mu_{g+\alpha f}(fh|t) - \mu_{g+\alpha f}(f|t)\mu_{g+\alpha f}(h|t)\}. \tag{2.4}$$

By the mean value theorem,

$$\begin{aligned}
\text{SKL}(\hat{g}, g_0) &= \mu_{\hat{g}}(\hat{g} - g_0) - \mu_{g_0}(\hat{g} - g_0) \\
&= \int_{\mathcal{T}} m(t)\{\mu_{g_0+\alpha(\hat{g}-g_0)}((\hat{g} - g_0)^2|t) - (\mu_{g_0+\alpha(\hat{g}-g_0)}(\hat{g} - g_0|t))^2\} \\
&\approx \int_{\mathcal{T}} m(t)\{\mu_{g_0}((\hat{g} - g_0)^2|t) - (\mu_{g_0}(\hat{g} - g_0|t))^2\} \\
&= V(\hat{g} - g_0),
\end{aligned}$$

where $\alpha \in [0, 1]$ and

$$V(h) = V(h, h) = V_{g_0}(h, h) = \int_{\mathcal{T}} m(t)v_{g_0}(h|t) = \int_{\mathcal{T}} m(t)\{\mu_{g_0}(h^2|t) - (\mu_{g_0}(h|t))^2\}, \tag{2.5}$$

and $v_g(h|t) = v_g(h, h|t) = \mu_g(h^2|t) - (\mu_g(h|t))^2$. $V(\hat{g} - g_0)$ is a properly weighted mean square error.

Under the condition $\mathcal{X} = \overline{\cup_{t:m(t)>0}\{x : w(t,x) > 0\}}$, $V(g)$ defines a square norm in $\mathcal{H} \not\supset \text{span}\{1\}$ which is of direct interest under the stochastic structure. $J(g)$ defines the notion of smoothness. The characterization of smoothness 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 $B(\eta) \leq \epsilon A(\eta)$; see Weinberger (1974, §3.3).

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

Under A.1, using Theorem 3.1 of Weinberger (1974, p.52), 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; see Gu and Qiu (1991, §4). Fourier expansion $g = \sum_\nu g_\nu \phi_\nu$ also exists for $g \in \mathcal{H}$ when convergence is defined in the $(V + J)$ norm. 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 (2.2) for any fixed λ restricts the model space to an effectively finite dimension in terms of the V norm, which is necessary for any possible 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 smoothness of members of $\mathcal{H}$.

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

The asymptotic behavior of the estimator depends on n, λ, and r.

3 Asymptotic Convergence

Assume $g_0 \in \mathcal{H}$. Let g_1 be the minimizer of

$$Q_\lambda(g) = -\frac{1}{n}\sum_{i=1}^n g(X_i) + \frac{1}{n}\sum_{i=1}^n \mu_{g_0}(g|t_i) + \frac{1}{2}V(g - g_0) + \frac{\lambda}{2}J(g). \tag{3.1}$$

Substituting in the Fourier expansions $g = \sum_\nu g_\nu \phi_\nu$ and $g_0 = \sum_\nu g_{\nu,0}\phi_\nu$, the Fourier coefficients of g_1 can be easily solved to be $g_{\nu,1} = (\beta_\nu + g_{\nu,0})/(1 + \lambda\rho_\nu)$, where $\beta_\nu = n^{-1}\sum_{i=1}^n (\phi_\nu(X_i) - \mu_{g_0}(\phi_\nu|t_i))$. It is obvious that $E\beta_\nu = 0$ and $E\beta_\nu^2 = n^{-1}$. Theorem 4.1 of Gu and Qiu (1991) holds verbatim but with the modified definitions of V and g_1 in the current more general setup.

Theorem 3.1 *Under A.1 and A.2, as $n \to \infty$ and $\lambda \to 0$, $V(g_1 - g_0) = O_p(n^{-1}\lambda^{-1/r} + \lambda)$ and $\lambda J(g_1 - g_0) = O_p(n^{-1}\lambda^{-1/r} + \lambda)$.*

Let $A_{g,h}(\alpha) = P_\lambda(g + \alpha h)$ and $B_{g,h}(\alpha) = Q_\lambda(g + \alpha h)$. It is clear that

$$0 = \dot{A}_{\hat{g},\hat{g}-g_1}(0) = -\frac{1}{n}\sum_{i=1}^{n}(\hat{g} - g_1)(X_i) + \frac{1}{n}\sum_{i=1}^{n}\mu_{\hat{g}}(\hat{g} - g_1|t_i) + \lambda J(\hat{g}, \hat{g} - g_1) \tag{3.2}$$

and

$$0 = \dot{B}_{g_1,\hat{g}-g_1}(0) = -\frac{1}{n}\sum_{i=1}^{n}(\hat{g}-g_1)(X_i)+\frac{1}{n}\sum_{i=1}^{n}\mu_{g_0}(\hat{g}-g_1|t_i)+V(g_1-g_0,\hat{g}-g_1)+\lambda J(g_1,\hat{g}-g_1). \tag{3.3}$$

Subtracting (3.3) from (3.2), some algebra yields

$$\frac{1}{n}\sum_{i=1}^{n}\{\mu_{\hat{g}}(\hat{g}-g_1|t_i)-\mu_{g_1}(\hat{g}-g_1|t_i)\}+\lambda J(\hat{g}-g_1) = V(g_1-g_0,\hat{g}-g_1)-\frac{1}{n}\sum_{i=1}^{n}\{\mu_{g_1}(\hat{g}-g_1|t_i)-\mu_{g_0}(\hat{g}-g_1|t_i)\}. \tag{3.4}$$

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

$$\exists c_1, c_2 \in (0,\infty) \text{ such that } c_1 v_{g_0}(h|t) \le v_g(h|t) \le c_2 v_{g_0}(h|t),\ \forall t \in \mathcal{T}.$$

A.3 assures the equivalence of the V distance and the SKL in B_0.

Assumption A.4. $\exists c_3 < \infty$ such that $\int_{\mathcal{T}} m(t)(v_{g_0}(\phi_\nu|t))^2 \le c_3,\ \forall \nu$.

A.4 is trivial when $\mathcal{T}$ is finite, noting that $V(\phi_\nu) = 1$.

Theorem 3.2 *Under A.1 - A.4, as* $\lambda \to 0$ *and* $n\lambda^{2/r} \to \infty$, $V(\hat{g} - g_1) = O_p(n^{-1}\lambda^{-1/r} + \lambda)$ *and* $\lambda J(\hat{g} - g_1) = O_p(n^{-1}\lambda^{-1/r} + \lambda)$. *Consequently,* $V(\hat{g} - g_0) = O_p(n^{-1}\lambda^{-1/r} + \lambda)$, $\lambda J(\hat{g} - g_0) = O_p(n^{-1}\lambda^{-1/r} + \lambda)$, *and* $\mathrm{SKL}(\hat{g}, g_0) = O_p(n^{-1}\lambda^{-1/r} + \lambda)$.

Proof: The second part of the theorem follows from the first part, Theorem 3.1, and A.3. Below is a proof of the first part. By A.3 and the mean value theorem,

$$c_1\frac{1}{n}\sum_{i=1}^{n}v_{g_0}(\hat{g} - g_1|t_i) \le \frac{1}{n}\sum_{i=1}^{n}\{\mu_{\hat{g}}(\hat{g} - g_1|t_i) - \mu_{g_1}(\hat{g} - g_1|t_i)\}. \tag{3.5}$$

Via the Fourier expansion $(\hat{g} - g_1) = \sum_\nu(\hat{g}_\nu - g_{\nu,1})\phi_\nu$,

$$\begin{aligned}
&|\frac{1}{n}\sum_{i=1}^{n}v_{g_0}(\hat{g} - g_1|t_i) - V(\hat{g} - g_1)| \\
= \ &|\sum_\nu\sum_\mu(\hat{g}_\nu - g_{\nu,1})(\hat{g}_\mu - g_{\mu,1})\{\frac{1}{n}\sum_{i=1}^{n}v_{g_0}(\phi_\nu,\phi_\mu|t_i) - V(\phi_\nu,\phi_\mu)\}| \\
\le \ &[\sum_\nu\sum_\mu(1+\lambda\rho_\nu)(1+\lambda\rho_\mu)(\hat{g}_\nu - g_{\nu,1})^2(\hat{g}_\mu - g_{\mu,1})^2]^{1/2}
\end{aligned}$$

$$
\begin{aligned}
& [\sum_{\nu}\sum_{\mu}(1+\lambda\rho_\nu)^{-1}(1+\lambda\rho_\mu)^{-1}\{\frac{1}{n}\sum_{i=1}^{n} v_{g_0}(\phi_\nu,\phi_\mu|t_i)-V(\phi_\nu,\phi_\mu)\}^2]^{1/2} \\
= \;& (V+\lambda J)(\hat{g}-g_1)O_p(n^{-1/2}\lambda^{-1/r}) \\
= \;& (V+\lambda J)(\hat{g}-g_1)o_p(1), \qquad (3.6)
\end{aligned}
$$

where $\sum_\nu(1+\lambda\rho_\nu)^{-1} = O(\lambda^{-1/r})$ (Gu and Qiu, 1991, Lemma 4.2) and $\{n^{-1}\sum_{i=1}^n v_{g_0}(\phi_\nu,\phi_\mu|t_i) - V(\phi_\nu,\phi_\mu)\}^2 = O_p(n^{-1})$ (by A.4). Similarly,

$$
\frac{1}{n}\sum_{i=1}^{n}\{\mu_{g_1}(\hat{g}-g_1|t_i)-\mu_{g_0}(\hat{g}-g_1|t_i)\} = c\frac{1}{n}\sum_{i=1}^{n} v_{g_0}(g_1-g_0,\hat{g}-g_1|t_i), \qquad (3.7)
$$

where $c \in [c_1, c_2]$, and

$$
|\frac{1}{n}\sum_{i=1}^{n} v_{g_0}(g_1-g_0,\hat{g}-g_1|t_i)-V(g_1-g_0,\hat{g}-g_1)| = (V+\lambda J)^{1/2}(\hat{g}-g_1)(V+\lambda J)^{1/2}(g_1-g_0)o_p(1). \quad (3.8)
$$

Combining (3.5) through (3.8), (3.4) leads to

$$
(c_1V+\lambda J)(\hat{g}-g_1)(1+o_p(1)) \le (V+\lambda J)^{1/2}(\hat{g}-g_1)(V+\lambda J)^{1/2}(g_1-g_0)(|c-1|+o_p(1)). \quad (3.9)
$$

The first part of the theorem follows from (3.9) and Theorem 3.1. □

Note that for a singleton $\mathcal{T}$, A.3 can be reduced to $c_1 v_{g_0}(h|t) \le v_g(h|t)$ only, $n\lambda^{2/r} \to \infty$ to $n\lambda^{1/r} \to \infty$, yet the first part of Theorem 3.2 refined to $(V+\lambda J)(\hat{g}-g_1) = o_p(n^{-1}\lambda^{-1/r}+\lambda)$; $\mathrm{SKL}(\hat{g}, g_0) = O_p(n^{-1}\lambda^{-1/r}+\lambda)$ then needs a separate proof under the reduced A.3. See Gu and Qiu (1991, §4).

4 Semiparametric Adaptive Estimator

The space $\mathcal{H}$ is in general infinite dimensional and $\hat{g}$ not computable. I shall propose and justify a computable semiparametric adaptive estimator in this section. Given a norm in $J_\perp$, $\mathcal{H}$ has a tensor sum decomposition such that J is a square norm in $\mathcal{H} \ominus J_\perp$. Let $\mathcal{H}_n = J_\perp \oplus \mathrm{span}\{R_J(X_i,\cdot), i = 1,\cdots,n\}$ where R_J is the reproducing kernel of $(\mathcal{H}\ominus J_\perp, J)$. Define $\hat{g}_n$ to be the minimizer of (2.2) in $\mathcal{H}_n$.

Assumption A.5. $\int_{\mathcal{T}} m(t)\{v_{g_0}(\phi_\nu\phi_\mu|t) + (\mu_{g_0}(\phi_\nu\phi_\mu|t) - \mu_{g_0}(\phi_\nu\phi_\mu))^2\} \le c_4 < \infty, \forall \nu,\mu.$

Lemma 4.1 *Under A.1, A.2, A.4 and A.5, as* $\lambda \to 0$ *and* $n\lambda^{2/r} \to \infty$, $V(h) = \lambda J(h)o_p(1)$, $\forall h \in \mathcal{H}\ominus\mathcal{H}_n$.

Proof: For $h \in \mathcal{H} \ominus \mathcal{H}_n \subset \mathcal{H} \ominus J_\perp$, $h(X_i) = J(R_J(X_i, \cdot), h) = 0$. Similar to (3.6),

$$V(h) \le \mu_{g_0}(h^2) = \sum_\nu \sum_\mu h_\nu h_\mu \{\mu_{g_0}(\phi_\nu \phi_\mu) - \frac{1}{n}\sum_{i=1}^n \phi_\nu(X_i)\phi_\mu(X_i)\} \le (V + \lambda J)(h) o_p(1),$$

where $h = \sum_\nu h_\nu \phi_\nu$ is the Fourier expansion and $\{n^{-1}\sum_{i=1}^n \phi_\nu(X_i)\phi_\mu(X_i) - \mu_{g_0}(\phi_\nu\phi_\mu)\}^2 = O_p(n^{-1})$ by A.5. □

Theorem 4.1 *Let g_n be the projection of $\hat{g}$ in $\mathcal{H}_n$. Under A.1 - A.5, as $\lambda \to 0$ and $n\lambda^{2/r} \to \infty$, $V(\hat{g} - g_n) = o_p(n^{-1}\lambda^{-1/r} + \lambda)$ and $\lambda J(\hat{g} - g_n) = O_p(n^{-1}\lambda^{-1/r} + \lambda)$.*

Proof: From $\dot{A}_{\hat{g}, \hat{g} - g_n}(0) = 0$ (cf. (3.2)) and $J(g_n, \hat{g} - g_n) = 0$,

$$\lambda J(\hat{g} - g_n) = \frac{1}{n}\sum_{i=1}^n \{(\hat{g} - g_n)(X_i) - \mu_{g_0}(\hat{g} - g_n | t_i)\} + \frac{1}{n}\sum_{i=1}^n \{\mu_{g_0}(\hat{g} - g_n | t_i) - \mu_{\hat{g}}(\hat{g} - g_n | t_i)\}. \quad (4.1)$$

Using the technique of (3.6),

$$|\frac{1}{n}\sum_{i=1}^n \{(\hat{g} - g_n)(X_i) - \mu_{g_0}(\hat{g} - g_n | t_i)\}| = |\sum_\nu (\hat{g}_\nu - g_{\nu,n})\beta_\nu| \le (V + \lambda J)^{1/2}(\hat{g} - g_n) O_p(n^{-1/2}\lambda^{-1/2r}). \quad (4.2)$$

Similar to (3.7) and (3.8),

$$|\frac{1}{n}\sum_{i=1}^n \{\mu_{g_0}(\hat{g} - g_n | t_i) - \mu_{\hat{g}}(\hat{g} - g_n | t_i)\}| = cV(\hat{g} - g_n, \hat{g} - g_0) + (V + \lambda J)^{1/2}(\hat{g} - g_n)(V + \lambda J)^{1/2}(\hat{g} - g_0) o_p(1). \quad (4.3)$$

Combining (4.1) - (4.3) and Theorem 3.2,

$$\lambda J(\hat{g} - g_n) \le (V + \lambda J)^{1/2}(\hat{g} - g_n) O_p(n^{-1/2}\lambda^{-1/2r} + \lambda^{1/2}). \quad (4.4)$$

The theorem follows from (4.4) and Lemma 4.1. □

Theorem 4.2 *Modify A.3 to also include g_n and $\hat{g}_n$ in the convex set B_0. Under A.1 - A.5, as $\lambda \to 0$ and $n\lambda^{2/r} \to \infty$, $V(\hat{g}_n - g_n) = O_p(n^{-1}\lambda^{-1/r} + \lambda)$ and $\lambda J(\hat{g}_n - g_n) = O_p(n^{-1}\lambda^{-1/r} + \lambda)$. Consequently, $V(\hat{g}_n - g_0) = O_p(n^{-1}\lambda^{-1/r} + \lambda)$, $\lambda J(\hat{g}_n - g_0) = O_p(n^{-1}\lambda^{-1/r} + \lambda)$, and $\mathrm{SKL}(\hat{g}_n, g_0) = O_p(n^{-1}\lambda^{-1/r} + \lambda)$.*

Proof: It suffices to prove the first part. From $\dot{A}_{\hat{g}_n, \hat{g}_n - g_n}(0) = \dot{A}_{\hat{g}, \hat{g}_n - \hat{g}}(0) = 0$ (cf. (3.2)), noting that $J(\hat{g} - g_n, g_n) = J(\hat{g} - g_n, \hat{g}_n) = 0$ so $J(\hat{g}, \hat{g}_n - \hat{g}) = J(g_n, \hat{g}_n - g_n) - J(\hat{g} - g_n)$, it can be shown

that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\{\mu_{\hat{g}_n}(\hat{g}_n - g_n|t_i) - \mu_{g_n}(\hat{g}_n - g_n|t_i)\} + \lambda J(\hat{g}_n - g_n) + \lambda J(\hat{g} - g_n) \\
= \;& \frac{1}{n}\sum_{i=1}^{n}\{\mu_{\hat{g}}(\hat{g}_n - g_n|t_i) - \mu_{g_n}(\hat{g}_n - g_n|t_i)\} + \frac{1}{n}\sum_{i=1}^{n}\{\mu_{\hat{g}}(g_n - \hat{g}|t_i) - \mu_{g_0}(g_n - \hat{g}|t_i)\} \\
&+\frac{1}{n}\sum_{i=1}^{n}\{\mu_{g_0}(g_n - \hat{g}|t_i) - (g_n - \hat{g})(X_i)\}. \qquad (4.5)
\end{aligned}$$

Similar to (3.5) and (3.6),

$$c_1 V(\hat{g}_n - g_n) + (V + \lambda J)(\hat{g}_n - g_n)o_p(1) \le \frac{1}{n}\sum_{i=1}^{n}\{\mu_{\hat{g}_n}(\hat{g}_n - g_n|t_i) - \mu_{g_n}(\hat{g}_n - g_n|t_i)\}. \qquad (4.6)$$

Noting that $V(\hat{g} - g_n) = \lambda J(\hat{g} - g_n)o_p(1)$,

$$|\frac{1}{n}\sum_{i=1}^{n}\{\mu_{\hat{g}}(\hat{g}_n - g_n|t_i) - \mu_{g_n}(\hat{g}_n - g_n|t_i)\}| = (V + \lambda J)^{1/2}(\hat{g}_n - g_n)(\lambda J)^{1/2}(\hat{g} - g_n)o_p(1) \qquad (4.7)$$

and

$$|\frac{1}{n}\sum_{i=1}^{n}\{\mu_{\hat{g}}(g_n - \hat{g}|t_i) - \mu_{g_0}(g_n - \hat{g}|t_i)\}| = (V + \lambda J)^{1/2}(\hat{g} - g_0)(\lambda J)^{1/2}(\hat{g} - g_n)o_p(1). \qquad (4.8)$$

Combining (4.6) through (4.8) and (4.2), (4.5) leads to

$$(c_1 V+\lambda J)(\hat{g}_n-g_n)(1+o_p(1)) \le (V+\lambda J)^{1/2}(\hat{g}_n-g_n)o_p(n^{-1/2}\lambda^{-1/2r}+\lambda^{1/2})+O_p(n^{-1}\lambda^{-1/r}+\lambda). \qquad (4.9)$$

The first part of the theorem follows. □

For a singleton $\mathcal{T}$, Theorem 4.2 remains valid when A.3 is reduced to $c_1 v_{g_0}(h|t) \le v_g(h|t)$ only, but $\mathrm{SKL}(\hat{g}_n, g_0) = O_p(n^{-1}\lambda^{-1/r} + \lambda)$ needs a separate proof; see Gu and Qiu (1991, §5).

5 Computation

Let $\{\phi_\nu, \nu = 1, \cdots, M\}$ span $J_\perp$ and $\xi_i = R_J(X_i, \cdot)$. By definition, a function in $\mathcal{H}_n$ has an expression

$$g = \sum_{i=1}^{n} c_i\xi_i + \sum_{\nu=1}^{M} d_\nu\phi_\nu = \boldsymbol{\xi}^T\boldsymbol{c} + \boldsymbol{\phi}^T\boldsymbol{d}, \qquad (5.1)$$

where $\boldsymbol{\xi}$ and $\boldsymbol{\phi}$ are vectors of functions and $\boldsymbol{c}$ and $\boldsymbol{d}$ are vectors of coefficients. Substituting (5.1) into (2.2), $\hat{g}_n$ can be calculated via minimizing

$$P_\lambda(\boldsymbol{c}, \boldsymbol{d}) = -\frac{1}{n}\mathbf{1}^T(Q\boldsymbol{c} + S\boldsymbol{d}) + \frac{1}{n}\sum_{i=1}^{n}\log\int_{\mathcal{X}} w(t_i, x)\exp\{\boldsymbol{\xi}^T\boldsymbol{c} + \boldsymbol{\phi}^T\boldsymbol{d}\} + \frac{\lambda}{2}\boldsymbol{c}^T Q\boldsymbol{c} \qquad (5.2)$$

with respect to $\boldsymbol{c}$ and $\boldsymbol{d}$, where Q is $n \times n$ with (i,j)th entry $\xi_i(X_j) = R_J(X_i, X_j)$ and S is $n \times M$ with (i,ν)th entry $\phi_\nu(X_i)$.

Let $\tilde{g} = \boldsymbol{\xi}^T\tilde{\boldsymbol{c}} + \boldsymbol{\phi}^T\tilde{\boldsymbol{d}}$ be the current estimate of g. For fixed λ, the one-step Newton update for minimizing (5.2) satisfies

$$\begin{pmatrix} V_{\xi,\xi} + \lambda Q & V_{\xi,\phi} \\ V_{\phi,\xi} & V_{\phi,\phi} \end{pmatrix} \begin{pmatrix} \boldsymbol{c} \\ \boldsymbol{d} \end{pmatrix} = \begin{pmatrix} Q\mathbf{1}/n - \mu_\xi + V_{\xi,g} \\ S^T\mathbf{1}/n - \mu_\phi + V_{\phi,g} \end{pmatrix}, \tag{5.3}$$

where $V_{\xi,\xi}$ is of size $n \times n$ with (i,j)th entry $n^{-1}\sum_{l=1}^n v_{\tilde{g}}(\xi_i, \xi_j|t_l)$, $V_{\xi,\phi}$ $n \times M$ with (i,ν)th entry $n^{-1}\sum_{l=1}^n v_{\tilde{g}}(\xi_i, \phi_\nu|t_l)$, $V_{\phi,\phi}$ $M \times M$ with (ν,μ)th entry $n^{-1}\sum_{l=1}^n v_{\tilde{g}}(\phi_\nu, \phi_\mu|t_l)$, $V_{\xi,g}$ $n \times 1$ with ith entry $n^{-1}\sum_{l=1}^n v_{\tilde{g}}(\xi_i, \tilde{g}|t_l)$, $V_{\phi,g}$ $M \times 1$ with νth entry $n^{-1}\sum_{l=1}^n v_{\tilde{g}}(\phi_\nu, \tilde{g}|t_l)$, μ_ξ $n \times 1$ with ith entry $n^{-1}\sum_{l=1}^n \mu_{\tilde{g}}(\xi_i|t_l)$, and μ_ϕ $M \times 1$ with νth entry $n^{-1}\sum_{l=1}^n \mu_{\tilde{g}}(\phi_\nu|t_l)$.

The choice of λ is crucial to the performance of the estimator. Among natural performance criteria are $\mathrm{SKL}(\hat{g}_n, g_0)$ and $V(\tilde{g}_n, g_0)$, where the mixing function $m(t)$ is to be substituted by the empirical distribution of t_i as in (5.3). From $\tilde{g}$, the one-step Newton update provides a group of estimates with a variable λ, and it is natural for one to try to calculate a better performing update by selecting a proper λ. Based on $\tilde{g}$, $L_{\tilde{g}}(g, g_0) = V_{\tilde{g}}(g)/2 - V_{\tilde{g}}(g, \tilde{g}) + \mu_{\tilde{g}}(g) - \mu_{g_0}(g)$ is a proxy of $\mathrm{SKL}(g, g_0)$ or $V(g - g_0)$, where $\mu_{g_0}(g)$ may be estimated using some variate of the sample mean and other terms can be calculated directly. A performance-oriented iteration can then be conducted to jointly update (λ, g) by choosing λ to minimize $\hat{L}_{\tilde{g}}(g, g_0)$ for g in the group of the one-step Newton updates, where $\hat{L}_{\tilde{g}}(g, g_0)$ is $L_{\tilde{g}}(g, g_0)$ with an estimated $\mu_{g_0}(g)$. Relevant discussions, formulas, and an algorithm for a singleton $\mathcal{T}$ can be found in Gu (1991, §§3-4), which hold verbatim in the more general setup of this article with the modified definitions of quantities in (5.3). Note that the algorithm does not require the paired data (t_i, X_i), but only the samples X_i and the empirical distribution of the biasing functions indexed by t_i.

6 Simulations

The simulations in this section augment the univariate simulations of Gu (1991, §5). The test density was chosen to be proportional to

$$f_0(x) = \frac{1}{3}e^{-50(x-.3)^2} + \frac{2}{3}e^{-50(x-.7)^2} \tag{6.1}$$

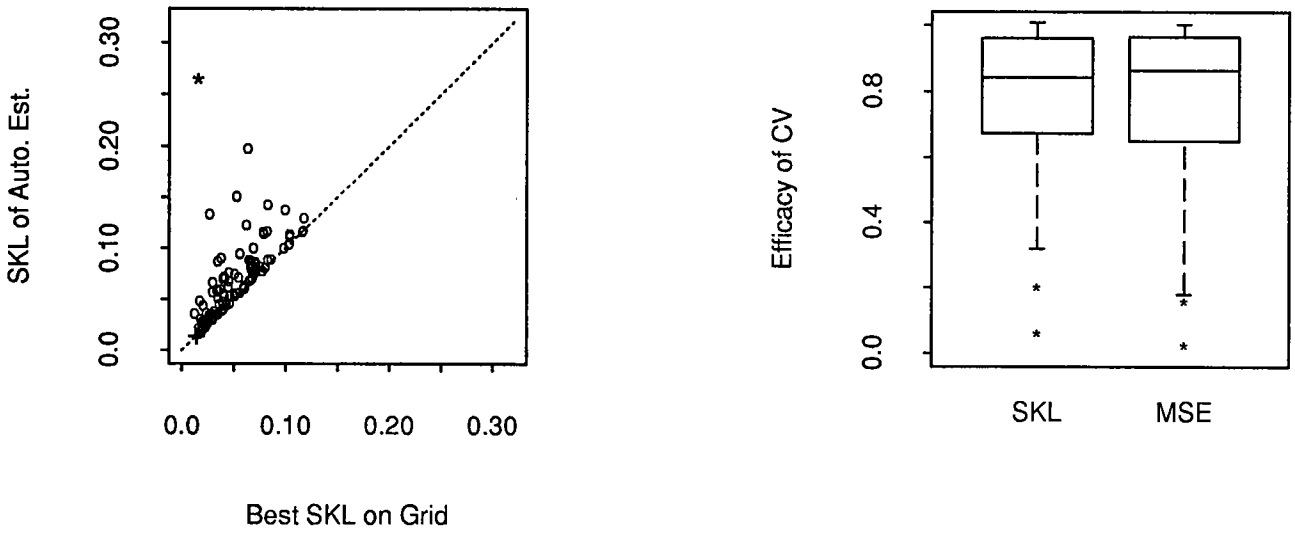
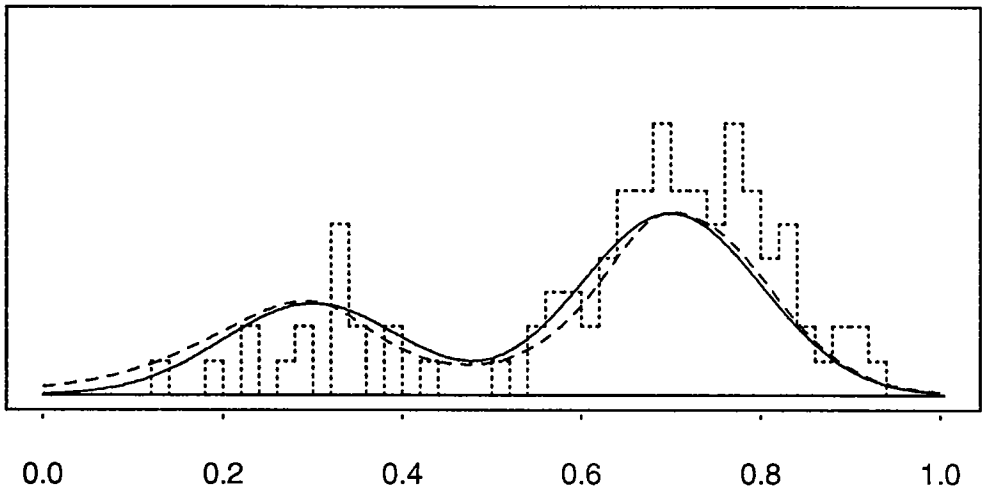


Figure 6.1: Efficacy of Automatic Algorithm on Length-Biased Data.

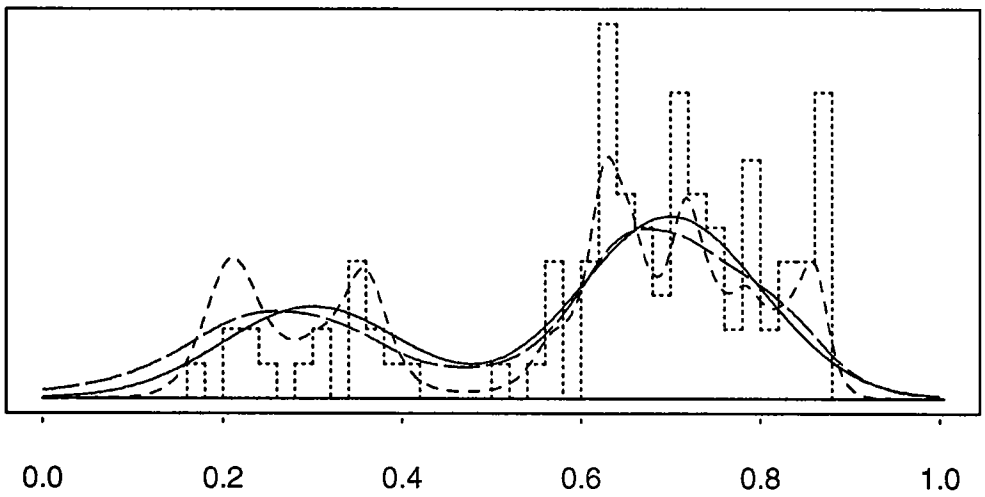
on $\mathcal{X} = [0,1]$. As specified in Gu (1991), $J_\perp = \{(\cdot - .5)\}$ and $R_J(x,y) = k_2(x)k_2(y) - k_4(|x-y|)$ which correspond to cubic spline smoothing, where $k_2 = (k_1^2 - 1/12)/2$, $k_4 = (k_1^4 - k_1^2/2 + 7/240)/24$, and $k_1 = (\cdot - .5)$. The simulations of Gu (1991) were on the ordinary samples of Example 1.1. The simulations presented here are on the length-biased samples of Example 1.2 and on the mixture of ordinary and length-biased samples of Example 1.3.

For Example 1.2, I generated 100 sets of length-biased data of sizes $n = 100$ from (6.1). The performance-oriented iteration converged on 99 data sets. Fixed-λ solutions of (5.2) were also calculated on a grid $\log_{10}\lambda = (-7)(.2)(-3)$ for all the data sets. In the computation, the integrals appearing in the quantities in (5.3) were approximated by summations over 300 equally spaced points on $(0,1)$; see Gu (1991, §4) for the motivation and justification of such a practice. The symmetrized Kullback-Leibler $\mathrm{SKL}(\hat{g}_n, g_0)$ and the mean square error (MSE) $V(\hat{g}_n - g_0)$ were calculated for all the automatic and fixed-λ estimates, where the integrals were also approximated by the summations over the 300 points.

For the 99 data sets on which the automatic algorithm converged, the minimum SKL and the minimum MSE of the fixed-λ estimates on the grid were identified. The SKL of the automatic estimate is plotted against that of the best estimate on the grid in the left frame of Figure 6.1.



An Ideal Automatic Fit



A Poor Automatic Fit

Figure 6.2: A Good Estimate and a Poor Estimate from Length-Biased Data.

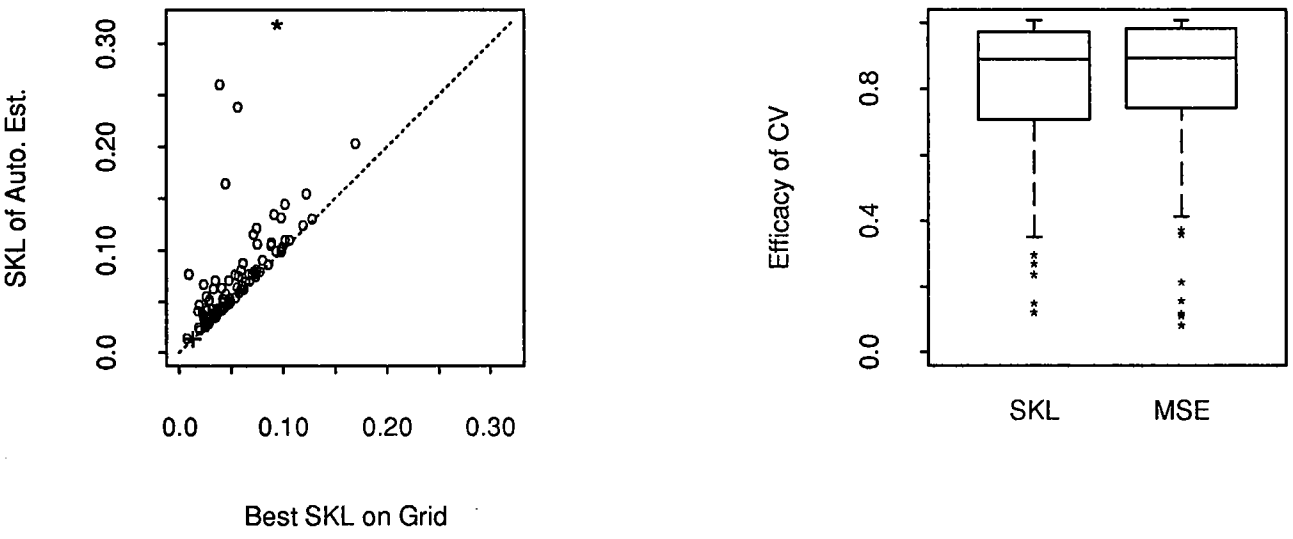
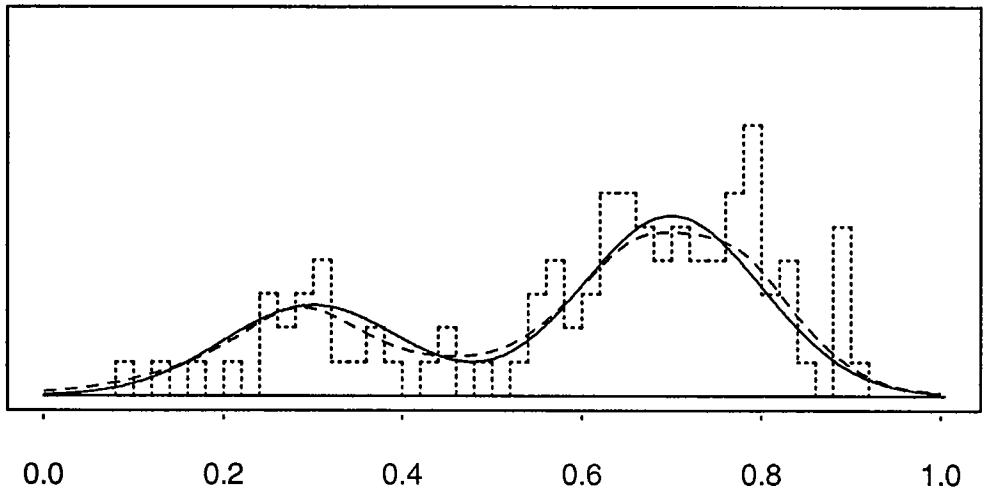


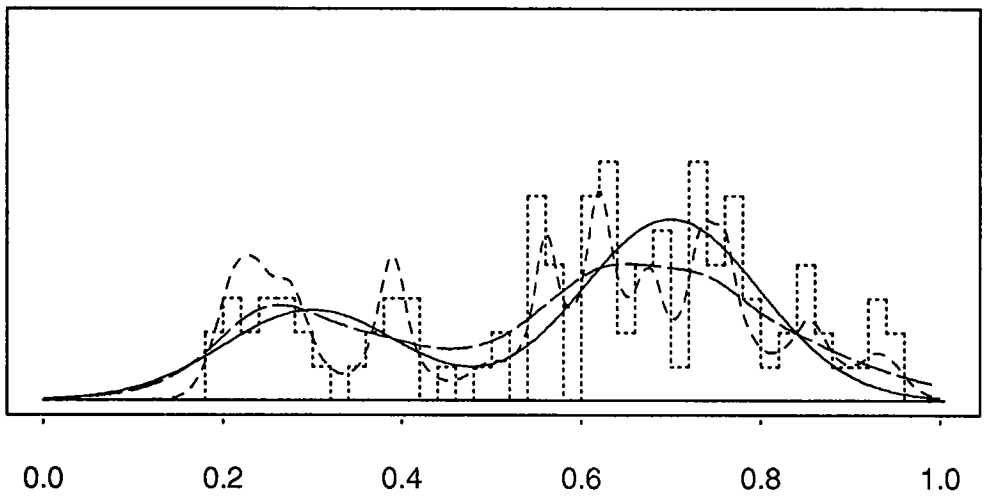
Figure 6.3: Efficacy of Automatic Algorithm on Mixture Data.

Two cases, the best and the worst performances of the automatic algorithm on the 99 data sets, are marked differently. A point on the dotted line indicates a perfect performance of the automatic algorithm. The efficacy of the automatic algorithm in SKL and MSE, defined by the ratio of the minimum score on the grid and the score of the automatic estimate, are summarized in the right frame of Figure 6.1 in box-plots. The best automatic estimate corresponding to the plus in the left frame of Figure 6.1 is plotted in the top frame of Figure 6.2 as the dashed line, superimposed with the true density as the solid line and the raw data as the finely-binned histogram in dotted lines. The worst automatic estimate corresponding to the star in the left frame of Figure 6.1 is similarly plotted in the bottom frame of Figure 6.2, where the best possible estimate on the grid is also superimposed as the dashed line with long dashes.

For Example 1.3, I generated 100 sets of mixed samples of sizes $n = 100$. In each of the data sets, 50 samples were drawn directly from (6.1) and 50 were drawn with a biasing function $w(x) = x$. The automatic algorithm converged on 99 data sets. The counter parts of Figures 6.1 and 6.2 are presented in Figures 6.3 and 6.4.



An Ideal Automatic Fit



A Poor Automatic Fit

Figure 6.4: A Good Estimate and a Poor Estimate from Mixture Data.

References

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

Cox, D. R. (1969). Some sampling problems in technology. In *New Developments in Survey Sampling.* Johnson, N. L. and Smith, H. Jr. (eds.), 506 – 527. Wiley, New York.

Gill, R. D., Vardi, Y., and Wellner, J. A. (1988). Large sample theory of empirical distributions in biased sampling models. *Ann. Statist.*. **16**, 1069 – 1112.

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

Gu, C. and Qiu, C. (1991), "Smoothing spline density estimation: Theory," Technical Report 91-19, Purdue University, Dept. Statistics.

Gu, C. (1991). Smoothing spline density estimation: A dimensionless automatic algorithm. Technical Report 91-41, Purdue University, Dept. Statistics.

Jones, M. C. (1991). Kernel density estimation for length biased data. *Biometrika*, **78**, 511 – 520.

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

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

Vardi, Y. (1982). Nonparametric estimation in the presence of length bias. *Ann. Statist.*, **10**, 616 – 620.

—— (1985). Empirical distributions in selection bias models. *Ann. Statist.*, **13**, 178 – 203.

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.