Spline Density Estimation and Inference with Model-Based Penalities

Jian Shi, Anna Liu and Yuedong Wang *

January 3, 2017

Abstract

In this paper we propose model-based penalties for smoothing spline density estimation and inference. These model-based penalties incorporate indefinite prior knowledge that the density is close to, but not necessarily in a family of distributions. We will use the Pearson and generalization of the generalized inverse Gaussian families to illustrate the derivation of penalties and reproducing kernels. We also propose new inference procedures to test the hypothesis that the density belongs to a specific family of distributions. We conduct extensive simulations to show that the model-based penalties can substantially reduce both bias and variance in the decomposition of the Kullback-Leibler distance, and the new inference procedures are more powerful than some existing ones.

*Jian Shi (email: shi@pstat.ucsb.edu) is PhD student, Department of Statistics and Applied Probability, University of California, Santa Barbara, California 93106. Anna Liu (email: anna@math.umass.edu) is Associate Professor, Department of Mathematics and Statistics, University of Massachusetts, Amherst, MA 01002. Yuedong Wang (email: yuedong@pstat.ucsb.edu) is Professor, Department of Statistics and Applied Probability, University of California, Santa Barbara, California 93106. Anna Liu's research was supported by a grant from the National Science Foundation (DMS-1507078). Yuedong Wang's research was supported by a grant from the National Science Foundation (DMS-1507620). Address for correspondence: Yuedong Wang, Department of Statistics and Applied Probability, University of California, Santa Barbara, California 93106.

KEY WORDS: Generalization of the generalized inverse Gaussian family, Goodness-of-fit test, L-spline, Pearson family, penalized likelihood.

1 Introduction

Density estimation has been widely studied due to its principal role in statistics and machine learning. Many methods such as kernel (Silverman 1986), local likelihood (Loader 1999) and smoothing spline (Gu 2013) have been developed to estimate density functions nonparametrically. These nonparametric techniques allow data to speak for themselves.

Often there is prior information suggesting that the density function can be well approximated by a parametric family of densities. For example, it may be known that the density is close to, but not necessarily is a Gamma distribution. This kind of indefinite information has not been explored in the field of density estimation. In his classic book on density estimation, Silverman (1984) alluded that different penalties may be considered for different situations in the context of penalized likelihood density estimation. In particular, he suggested penalties to the second and third derivatives of the logarithm of the density so that zero penalties correspond to the exponential and normal density functions respectively. To the best of our knowledge, no research has been done to incorporate indefinite prior information into the construction of the penalties.

We will consider different penalties through L-splines in this paper. The L-spline has been developed to incorporate prior knowledge in nonparametric regression models. It is known that the L-spline can reduce bias in the estimation of a regression function (Wahba 1990, Heckman & Ramsay 2000, Wang 2011, Gu 2013). The goal of this paper is to develop novel density estimation methods that can incorporate indefinite prior knowledge and consequently lead to better estimation procedures. In particular, we will consider model-based penalties for the Pearson family and the generalization of the generalized inverse Gaussian (GGIG) family, and derive penalties and

reproducing kernels for some special cases in these families of distributions. We will show that the model-based penalties can substantially reduce both bias and variance in the decomposition of the Kullback-Leibler (KL) distance of smoothing spline estimates of density functions. Many methods have been developed in the literature to test the hypothesis that the density belongs to a specific family of distributions (Anderson & Darling 1954, Stephens 1974, Stephens 1986). We will develop new inference procedures based on L-spline estimates. To the best of our knowledge, this paper is the first to employ L-splines for density estimation and inference.

The remainder of the article is organized as follows. Section 2 reviews smoothing spline density estimation and L-splines. Sections 3 and 4 present model constructions for the Pearson and GGIG families respectively. Section 5 introduces new inference procedures based on L-spline estimates. Section 6 presents simulation studies to compare the proposed L-spline based estimation and inference procedures with existing methods.

2 Smoothing Spline Density Estimation and L-Splines

2.1 Smoothing spline density estimation

Let $X_1, \ldots, X_n$ be independent and identically distributed (iid) random samples with a probability density $f(x)$ on an interval $[a, b]$. We assume that $f > 0$ on $[a, b]$. To enforce the conditions of $f > 0$ and $\int_a^b f = 1$ for a density function, throughout this article we will use the logistic transformation, $f = \exp(g)/\int_{\mathcal{X}} \exp(g)$, where g will be referred to as the logistic transformation of f (Gu 2013). We will model and estimate the function g which is free of constraints.

Assume that $g \in \mathcal{H}$ where $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS). To make the logistic transformation one-to-one, the constant functions are removed from the space $\mathcal{H}$ (Gu 2013). A

popular choice of $\mathcal{H}$ is $\mathcal{H} = W_2^m[a,b] \ominus \{1\}$ where

$$W_2^m[a,b] = \{f : f, f', \ldots, f^{(m-1)} \text{are absolutely continuous}, \int_a^b (f^{(m)})^2 dx < \infty\} \tag{1}$$

is the Sobolev space. In this article we assume that $g \in W_{20}^m[a,b]$ where $W_{20}^m[a,b] = W_2^m[a,b] \ominus \{1\}$.

A smoothing spline estimate of g is the minimizer of the penalized likelihood

$$-\frac{1}{n}\sum_{i=1}^{n} g(X_i) + \log \int_a^b e^g dx + \frac{\lambda}{2} J(g), \tag{2}$$

in $\mathcal{H}$, where $J(g)$ is a square (semi) norm penalty. The solution to (2) does not fall in a finite dimensional space. Let $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$ where $\mathcal{H}_0 = \{g : \ J(g) = 0\}$ and $\mathcal{H}_1$ is an RKHS with reproducing kernel (RK) R_1. Consider the finite-dimentional space $\mathcal{H}^* = \mathcal{H}_0 \oplus \text{span}\{R_1(Z_j, .),\ j = 1, ..., q\}$ where $\{Z_j\}$ is a random subset of $\{X_i\}$. As in Gu & Wang (2003) we will approximate the solution to (2) by the solution in the finite dimensional space $\mathcal{H}^*$. Gu & Wang (2003) showed that with appropriate choice of q, the approximation is efficient in the sense that the estimates in the whole model space $\mathcal{H}$ and the subspace $\mathcal{H}^*$ have the same convergence rate. Asymptotic properties were studied by Gu & Qiu (1993). We will use the approximate cross-validation estimate of the relative KL distance to select the smoothing parameter λ. See Gu (2013) for details.

2.2 Model-based penalty and L-splines

As discussed in Section 2.1, in the construction of a smoothing spline model, one needs to decide the penalty functional $J(g)$, or equivalently, the null space $\mathcal{H}_0$ consisting of functions which are not penalized. The most popular choice of the penalty is the roughness penalty with $J(g) = \int_a^b (g^{(m)})^2 dx$. When $m = 2$ and $m = 3$ respectively, the null spaces $\mathcal{H}_0$ are the linear and quadratic functions which correspond to the exponential and normal distributions suggested in Silverman (1984).

Often there exists information suggesting that f can be well approximated by a parametric fam-

ily of densities, and logistic transformation of density functions in this family satisfy the differential equation $Lg = 0$ where

$$L = D^m + \sum_{j=1}^{m-1} \omega_j(x) D^j \tag{3}$$

is a linear differential operator with $m \geq 1$, D^j is the jth derivative operator, and ω_i are continuous real-valued functions. Two such families of distributions, Pearson and GGIG, will be discussed in Sections 3 and 4.

An L-spline density estimate is the solution to (2) with penalty $J(g) = \int_a^b (Lg)^2 dx$. Instead of the standard roughness penalty, an L-spline uses a penalty constructed based on a parametric model. The null space $\mathcal{H}_0$ corresponds to the specified parametric family of densities. Therefore, it allows us to incorporate the information that g is close to, but not necessarily in the null space $\mathcal{H}_0$. Heckman & Ramsay (2000) called $\mathcal{H}_0$ as the favored parametric model. We will show in Section 6 that the model-based penalty can lead to better estimates of density functions. We will construct test procedures for the hypothesis that the density belongs to the specific parametric family in Section 5.

Since $g \in W_{20}^m[a, b]$, Lg exists and is square integrable. There exists real-valued functions, $\phi_1, ..., \phi_m$, such that they form a basis of $\mathcal{H}_0 = \{g : Lg = 0\}$. Let

$$W(x) = \begin{pmatrix} \phi_1(x) & \phi_2(x) & \cdots & \phi_m(x) \\ \phi_1'(x) & \phi_2'(x) & \cdots & \phi_m'(x) \\ \vdots & \vdots & & \vdots \\ \phi_1^{(m-1)}(x) & \phi_2^{(m-1)}(x) & \cdots & \phi_m^{(m-1)}(x) \end{pmatrix}$$

be the Wronskian matrix associated with $\phi_1, ..., \phi_m$, and

$$G(x, s) = \begin{cases} \boldsymbol{\phi}^T(x)\boldsymbol{\phi}^*(s), & s \leq x, \\ 0, & s > x, \end{cases}$$

be the Green function associated with L where $\boldsymbol{\phi}(x) = (\phi_1(x), \ldots, \phi_m(x))^T$ and $\boldsymbol{\phi}^*(x) = (\phi_1^*(x), \ldots, \phi_m^*(x))^T$ is the last column of $W^{-1}(x)$. Then $W_{20}^m[a,b]$ is an RKHS under the inner product

$$(f,g) = \sum_{\nu=0}^{m-1} f^{(\nu)}(a)g^{(\nu)}(a) + \int_a^b (Lf)(Lg)dx,$$

and $W_{20}^m[a,b] = \mathcal{H}_0 \oplus \mathcal{H}_1$, where $\mathcal{H}_0 =$ span $\{\phi_1, \ldots, \phi_m\}$ and $\mathcal{H}_1 = \{f \in W_{20}^m[a,b] : f^{(\nu)}(a) = 0, \nu = 0, \ldots, m-1\}$ are RKHS's with corresponding RKs

$$R_0(x,z) = \boldsymbol{\phi}^T(x)\{W^T(a)W(a)\}^{-1}\boldsymbol{\phi}(z), \tag{4}$$

$$R_1(x,z) = \int_a^b G(x,s)G(z,s)ds. \tag{5}$$

See Wang (2011) for details.

3 L-spline for Pearson Family of Distributions

The Pearson family is a continuous distribution system proposed by Karl Pearson (Pearson 1894). A Pearson density function $f(x)$ is any valid solution to the Pearson differential equation

$$\frac{1}{f(x)}\frac{df(x)}{dx} + \frac{a_0 + (x - a_5)}{a_1(x-a_5)^2 + a_2(x-a_5) + a_3} = 0, \tag{6}$$

where $a_0 = a_2 = \sqrt{\mu_2\beta_1}(\beta_2 + 3)/(10\beta_2 - 12\beta_1 - 18)$, $a_1 = (2\beta_2 - 3\beta_1 - 6)/(10\beta_2 - 12\beta_1 - 18)$, $a_3 = \mu_2(4\beta_2 - 3\beta_1)/(10\beta_2 - 12\beta_1 - 18)$, β_1 is the skewness, β_2 is the kurtosis, and μ_2 is the second central moments. Pearson identified 12 types of distributions based on different values of parameters. The Pearson family includes most commonly used distributions such as the uniform, exponential, normal, Gamma, Beta, inverse Gamma, Student's t and Cauchy distributions.

It is not difficult to show that the logistic transformation of density function in the Pearson family satisfy the differential equation $Lg = 0$ where

$$L = D^3 + \frac{2(2a_1(x-a_5) + a_2)}{a_1(x-a_5)^2 + a_2(x-a_5) + a_3}D^2 + \frac{2a_1}{a_1(x-a_5)^2 + a_2(x-a_5) + a_3}D. \tag{7}$$

Therefore we can construct model-based penalties using (7) for densities in the Pearson family. Explicit constructions can be derived for many special cases. We illustrate two such cases in the following two subsections.

3.1 Gamma distribution

The Gamma distribution (denoted as Gamma(α, β)) has density function

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \quad x > 0, \tag{8}$$

where $\alpha > 0$ and $\beta > 0$ are the shape and rate parameters, and Γ is the Gamma function.. It is a special case of the Pearson family (type III) with $a_1 = a_3 = a_5 = 0$, $a_2 = 1/\beta$ and $a_0 = -a_2(\alpha - 1)$. The logistic transformation of the density $g(x) = -\beta x + (\alpha - 1)\log(x)$.

Now consider the L-spline with model space $g(x) \in W^3_{20}[a, b]$ and differential operator

$$L = D^3 + \frac{2}{x} D^2. \tag{9}$$

As the domain of the Gamma distribution is $(0, \infty)$, we set a to be a small value closed to 0 and b large enough to cover all observations. The same method will be used for other distributions in the rest of this paper which are not defined on compact intervals. It can be shown that $\mathcal{H}_0 = \text{span}\{x, \log(x)\}$ and the RK of $\mathcal{H}_1$

$$\begin{aligned} R_1(x, z) \quad = \quad & [1 + \log(z) + \log(x) + \log(z)\log(x)] I_4(x \wedge z) - [z + x + z\log(x) + x\log(z)] I_3(x \wedge z) \\ & + xz I_2(x \wedge z) + I_{4,2}(x \wedge z) - [2 + \log(z) + \log(x)] I_{4,1}(x \wedge z) + (z + x) I_{3,1}(x \wedge z), \end{aligned}$$

where $x \wedge z = \min(x, z)$, $I_p(s) = \int_0^s x^p dx = s^{p+1}/(p+1)$, and $I_{p,k}(s) = \int_0^s x^p [\log(x)]^k = s^{p+1}[\log(s)]^k/(p+1) - k I_{p+1,k-1}(s)/(p+1)$. A brief derivation of the RK can be found in Appendix A.

3.2 Beta distribution

The Beta distribution has the density function

$$f(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 \le x \le 1, \tag{10}$$

where $\alpha > 0$ and $\beta > 0$ are the shape parameters. It is a special case of the Pearson family (type I) with $a_5 = a_3 = 0$, $a_1 = -a_2$, $\alpha = a_0/a_1 + 1$ and $\beta = (a_0 - 1)/a_1 + 1$. The logistic transformation $g(x) = (\alpha - 1)\log(x) + (\beta - 1)\log(1 - x)$.

Now consider the L-spline with model space $g(x) \in W^3_{20}[a, b]$ and differential operator

$$L = D^3 + \frac{2(2x-1)}{x(x-1)}D^2 + \frac{2}{x(x-1)}D. \tag{11}$$

It can be shown that $\mathcal{H}_0 = \text{span}\{\log(x), \log(1-x)\}$, and the RK of $\mathcal{H}_1$

$$\begin{aligned}
R_1(x,z) \;=\; & [\log(z)\log(1-x) + \log(x)\log(1-z)]I(x\wedge z;3,3,0,0) \\
& + \log(1-x)\log(1-z)I(x\wedge z;2,4,0,0) + \log(x)\log(z)I(x\wedge z;4,2,0,0) \\
& + I(x\wedge z;2,4,0,2) - (\log(x)+\log(z))I(x\wedge z;3,3,0,1) \\
& - [\log(1-x)+\log(1-z)]I(x\wedge z;2,4,0,1) \\
& + I(x\wedge z;4,2,2,0) - [\log(x)+\log(z)]I(x\wedge z;4,2,1,0) \\
& - [\log(1-x)+\log(1-z)]I(x\wedge z;3,3,1,0) + 2I(x\wedge z;3,3,1,1),
\end{aligned}$$

where

$$I(y;m_1,m_2,m_3,m_4) = \int_0^y x^{m_1}(1-x)^{m_2}\log(x)^{m_3}\log(1-x)^{m_4}dx.$$

A brief derivation of the RK is given in Appendix B.

4 L-spline for GGIG Family

Shakil, Kibria & Singh (2016) proposed the GGIG family of distributions to include some other commonly used distributions such as the inverse Gaussian, generlized inverse Gaussian (GIG),

Rayleigh and half-normal distributions which are not in the Pearson family. A GGIG density function $f(x)$ is the solution to the following differential equation

$$\frac{1}{f(x)}\frac{df(x)}{dx} = \frac{a_0 + a_p x^p + a_{2p}x^{2p}}{x^{p+1}}, \qquad x > 0. \tag{12}$$

The solution to the differential equation (12) is $f(x) = Cx^{\tau_1 - 1}\exp(-\tau_2 x^p - \tau_3 x^{-p})$ where $\tau_2 \geq 0$, $\tau_3 \geq 0$, $\tau_1 = a_p + 1$, $\tau_2 = -a_{2p}/p$, $\tau_3 = a_0/p$, and C is the normalizing constant. Then $g(x) = (\tau_1 - 1)\log(x) - \tau_2 x^p - \tau_3 x^{-p}$ which satisfies the differential equation $Lg = 0$ where

$$L = \sum_{k=0}^{p+1} \binom{2p+1}{k} \left(D^k x^{p+1}\right) D^{2p+2-k}. \tag{13}$$

The null space $\mathcal{H}_0 = \text{span}\{\log(x), x, \ldots, x^p, x^{-1}, \ldots, x^{-p}\}$.

We now consider the special case with $p = 1$ which includes many commonly used distributions such as the inverse Gaussian (IG) ($\tau_1 = -0.5$), GIG, reciprocal IG ($\tau_1 = 0.5$), hyperbolic ($\tau_1 = 1$), Gamma ($\tau_3 = 0$), inverse Gamma ($\tau_1 = 0$), Erlang ($\tau_1 > 0$ and is an integer, $\tau_3 = 0$), and exponential ($\tau_1 = 1$ and $\tau_3 = 0$). In this case we have $g(x) = (\tau_1 - 1)\log(x) - \tau_2 x - \tau_3 x^{-1}$ and

$$L = D^4 + 6x^{-1}D^3 + 6x^{-2}D^2. \tag{14}$$

It is not difficult to show that $\mathcal{H}_0 = \text{span}\{\log(x), x, x^{-1}\}$ and the RK of $\mathcal{H}_1$ is

$$\begin{aligned} R_1(x,z) \;=\; & \frac{1}{36xz}(x\wedge z)^9 - \frac{1}{16}(x\wedge z)^8 \log(x\wedge z)\left(\frac{1}{x}+\frac{1}{z}\right) \\ & + \frac{1}{16}(x\wedge z)^8\left(\frac{1}{8x}+\frac{1}{8z}+\frac{1}{z}\log(x)+\frac{1}{x}\log(z)\right) \\ & - \frac{1}{7}(x\wedge z)^7\log(x\wedge z)\left(\frac{2}{7}+\log(x)+\log(z)\right) + \frac{1}{7}(x\wedge z)^7\log(x\wedge z)^2 \\ & + \frac{1}{7}(x\wedge z)^7\left(\frac{2}{49}-\frac{z}{4x}-\frac{x}{4z}+\frac{1}{7}\log(x)+\frac{1}{7}\log(z)+\log(x)\log(z)\right) \\ & + \frac{1}{12}(x+z)(x\wedge z)^6\log(x\wedge z) - \frac{1}{12}(\frac{x}{6}+x\log(z)+\frac{z}{6}+z\log(x))(x\wedge z)^6 + \frac{1}{20}(x\wedge z)^5 xz. \end{aligned}$$

A brief derivation of the RK is given in Appendix C.

5 Inference of Density Using L-splines

Effective assessment of goodness-of-fit (GOF) and formal inference for a density function is critical in applications (Romantsova 1996, Del Castillo & Puig 1997, Lehmann & Romano 2006). In this section we consider the problem of deciding whether the density belongs to a parametric family of distributions. Let $X_1, \ldots, X_n$ be iid samples with a density $f(x)$ on an interval $[a, b]$. We consider the null hypothesis $H_0 : \ f \in \mathfrak{F}_0$ versus the alternative hypothesis $H_1 : \ f \notin \mathfrak{F}_0$ where $\mathfrak{F}_0$ is a specific family of distributions. We assume that there exists a differential operator L as in (3) such that $Lg = 0$ for all $f \in \mathfrak{F}_0$ where g is the logistic transformation f. Note that the null hypothesis H_0 is eqivalent to $g \in \mathcal{H}_0$.

5.1 Modified Anderson-Darling, Cramer-von Mises and Kolmogorov-Smirnov tests

A quadratic norm statistic based on the empirical distribution (EDF) is defined as

$$Q = n \int_{-\infty}^{\infty} (F_n(x) - F_0(x))^2 \, w(x) \, dF_0(x),$$

where F_n is the EDF, F_0 is an estimate of the cumulative density function (CDF) under the null hypothesis, and $w(x)$ is a weight function. Two well-konwn special cases are the Anderson-Darling (AD) and Cramer-von Mises (CVM) statistics with $w(x) = [F_0(x) \ (1 - F_0(x))]^{-1}$ and $w(x) = 1$ respectively (Stephens 1986).

Denote the CDF associated with the L-spline estimate of the density function as $F_s(x)$. Since L-splines with penalties constructed from specific families of distributions may provide better estimates of density functions (see Section 6), a natural extension of the AD and CVM statistics is to replace the EDF F_n in the quadratic norm statistic and weight function by F_s. The resulting modified testing methods are refered to as AD-L and CVM-L.

Kolmogorov-Smirnov test statistic is defined as

$$\text{KS} = \sup_x |F_n(x) - F_0|. \tag{15}$$

Again, we can construct a new test statistic by replacing F_n with $F_s(x)$. The resulting modified testing method is refered to as KS-L.

5.2 Likelihood ratio and Kullback-Leibler tests

The likelihood ratio (LR) statistic is

$$\text{LRT} = 2(l_s - l_0) \tag{16}$$

where l_s is the log-likelihood with the L-spline density estimate, and l_0 is the log-likelihood with MLE estimates of the parameters under the null hypothesis.

The KL distance between two density functions f_1 and f_2 is defined as

$$\text{KL}(f_1, f_2) = \int_a^b f_1(x) \log \frac{f_1(x)}{f_2(x)} dx. \tag{17}$$

Let f_0 be the estimated density under the null hypothesis, and f_s be the L-spline estimate of the density function. We will then use the KL distance between f_0 and f_s, $\text{KL}(f_0, f_s)$, as the KL test statistic.

6 Simulations

In this section, we conduct simulations to evaluate the proposed estimation and inference methods and compare them with existing methods. The function *ssden* in the R package *gss* is used to compute smoothing spline estimates of density functions (Gu 2013).

We will compare the estimation performance between the L-spline and cubic spline models. Denote f as the true density and $\hat{f}$ as an estimate. We will use the KL distance $\text{KL}(f, \hat{f})$ to assess

the performance of estimation. We will use the generalized decomposition

$$\mathrm{E}(\mathrm{KL}(f,\hat{f})) = \mathrm{KL}(f,\bar{f}) + \mathrm{E}(\mathrm{KL}(\bar{f},\hat{f})) = \text{bias } + \text{ variance} \tag{18}$$

proposed by Heskes (1998) to evaluate the bias-variance trade-off where $\bar{f} = \exp[\mathrm{E}(\log \hat{f})]/Z$ and Z is a normalization constant.

For density inference we will consider eight methods: Anderson-Darling (AD), Cramer-von Mises (CVM), Kolmogorov-Smirnov (KS), modified AD (AD-L), modified CVM (CVM-L), modified KS (KS-L), likelihood ratio (LR) and Kullback-Leibler (KL) tests. We will use the bootstrap method to approximate null distributions for all tests where the number of bootstrap samples is set to be 1000.

We will present results for two distributions, Gamma and inverse Guassian, as the favored parametric models. We will consider three sample size, $n = 100$, $n = 200$ and $n = 300$. Results for other distributions and sample sizes are similar. We generate 100 data replicates for each simulation setting.

6.1 Gamma distribution as the favored parametric model

The generalized Gamma family has the density function

$$f(x;\alpha,\beta,\delta) = \frac{\delta\beta^{\alpha}}{\Gamma(\alpha/\delta)} x^{\alpha-1} e^{-(\beta x)^{\delta}}, \qquad \alpha > 0,\ \beta > 0,\ \delta > 0, \quad x > 0. \tag{19}$$

The Gamma distribution Gamma(α,β) is a special case with $\delta = 1$. We set $\alpha = 2$ and $\beta = 1$ in our simulations, and consider three choices of δ: $\delta = 1$, $\delta = 2$, and $\delta = 3$ which reflect different degree of closeness to the Gamma distribution.

For each simulated data set, we compute the L-spline estimate of the density where L is given in (9) and the cubic spline estimate of the density. Table 1 lists biases, variances, and KL distances for the L-spline and cubic spline estimates under different simulation settings. The L-spline with

model-based penalty has smaller biases, variances, and KL distances than the cubic spline when $\delta = 1$ and $\delta = 2$. As expected, the improvement is larger when the true distribution is closer to the Gamma distribution.

δ	Model	n=100			n=200			n=300		
		Bias	Var	KL	Bias	Var	KL	Bias	Var	KL
1	Cubic	15.84	19.64	35.48	10.29	13.39	23.68	7.68	10.48	18.16
	L-spline	0.94	13.31	14.25	0.53	6.20	6.73	0.13	4.61	4.74
2	Cubic	7.61	15.07	22.68	5.98	9.30	15.28	4.14	7.07	11.21
	L-spline	1.78	16.05	17.83	0.92	9.76	10.67	0.89	6.22	7.11
3	Cubic	3.18	17.40	20.58	3.05	8.13	11.18	1.89	6.78	8.67
	L-spline	3.17	17.62	20.79	2.37	9.52	11.89	1.41	6.36	7.77

Table 1: Biases, variances, and KL distances in 10^{-3} with the generalized Gamma distribution.

For density inference we consider the null hypothesis that the distribution is Gamma. Table 2 lists powers of eights test methods with significance level set at 5%. The powers are the type I errors when $\delta = 1$. It is clear that all methods have type I errors smaller or close to 5%. With the EDF being replaced by the L-spline estimte, the modified AD, CVM and KS testsin general have larger powers than those from the original tests.

Table 3 lists more simulation results for testing the null hypothesis of a Gamma distribution against one of the distributions listed below:

1. The inverse gaussian distributionis defined in (21). We set $\kappa = 1$ and denote the density as $IG(\mu)$.

2. The lognormal distribution with density

$$f(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}}\, e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}\, \frac{1}{x\sigma\sqrt{2\pi}}\, e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}. \tag{20}$$

δ	Sample Size	AD	AD-L	CVM	CVM-L	KS	KS-L	LRT	KL
	100	0.19	0.10	0.18	0.04	0.14	0.16	0.15	**0.27**
0.6	200	0.37	0.29	0.34	0.19	0.31	0.36	0.32	**0.46**
	300	0.58	0.37	0.58	0.47	0.38	0.55	0.54	**0.64**
	100	0.06	0.06	0.05	0.05	0.05	0.04	0.06	0.05
1	200	0.04	0.05	0.06	0.04	0.05	0.04	0.05	0.05
	300	0.04	0.02	0.04	0.01	0.04	0.01	0.01	0.01
	100	0.13	**0.17**	0.15	0.16	**0.17**	0.14	**0.17**	0.14
2	200	0.38	**0.48**	0.30	0.47	0.24	0.46	0.45	0.42
	300	0.47	**0.56**	0.45	**0.56**	0.33	**0.56**	0.53	0.53
	100	0.25	**0.32**	0.22	0.31	0.17	**0.32**	0.31	0.31
3	200	0.48	**0.67**	0.39	0.64	0.32	0.64	0.66	0.66
	300	0.66	**0.77**	0.61	0.76	0.53	0.75	0.75	0.76

Table 2: Powers of eight test methods for the Gamma distribution.

We set $\mu = 0$, and denote the density as $LN(\sigma)$.

3. The Gompertz distribution with density

$$f(x;\eta,b) = b\eta e^{bx} e^{\eta} \exp\left(-\eta e^{bx}\right) b\eta e^{bx} e^{\eta} \exp\left(-\eta e^{bx}\right).$$

We set $b = 1$, and denote the density as $GO(1/\eta)$.

4. The linear failure rate distribution with density

$$f(x;\theta) = (1+\theta x)\exp\left(-x - \frac{\theta x^2}{2}\right).$$

We denote it as $LF(\theta)$.

We also calculate the skewness of each distribution. When the distributions under the alternative are GG(0.6,2), IG(1), IG(1.5) and LN(0.8) with which the skewness is greater than the Gamma distribution under the null (GG(1,2)), the KL statistic is more powerful. When the distributions under the alternative are GG(2,2), GG(3,2), GO(2), GO(4), LF(2) and LF(4) whose skewness is less than the Gamma distribution (GG(1,2)), the AD-L statistic is more powerful. We note that this pattern also holds for results listed in Table 2.

6.2 Inverse Gaussian distribution as the favored parametric model

The inverse Gaussian (IG) has density function

$$f(x;\mu,\kappa) = \Big(\frac{\kappa}{2\pi x^3}\Big)^{1/2} \exp\left\{\frac{-\kappa(x-\mu)^2}{2\mu^2 x}\right\}, \quad x>0, \tag{21}$$

where $\mu > 0$ is the mean and $\kappa > 0$ is the shape parameter. It belongs to the GGIG family with $p = 1$, $\tau_1 = -0.5$, $\tau_2 = 0.5\kappa/\mu^2$, and $\tau_3 = \kappa/2$. We set $\tau_2 = \tau_3 = 2$ in our simulations, and consider three choices of p: $p = 1$, $p = 2$ and $p = 3$ in (12), which reflect different degrees of closeness to the inverse Gaussian distribution.

For each simulated data set, we compute the L-spline estimate of the density where L is given in (14) and the cubic spline estimate of the density. Table 4 lists biases, variances, and KL distances for the L-spline and cubic spline estimates under different simulation settings. The L-spline with model-based penalty has smaller biases, variances, and KL distances than the cubic spline for all settings except when $p = 3$ and $n = 100$.

For density inference we consider the null hypothesis that the distribution is IG. We generate iid samples from the generlized inverse Gaussian (GIG) density

$$f(x) = \frac{(\alpha_0/\alpha_1)^{\zeta/2}}{2K_\zeta(\sqrt{\alpha_0\alpha_1})} x^{(\zeta-1)} e^{-\frac{(\alpha_0 x+\alpha_1/x)}{2}}, \quad \alpha_0>0,\ \alpha_1>0, x>0, \tag{22}$$

where K_ζ is a modified Bessel function of the second kind. The IG is a special case of GIG

Distribution	Size	AD	AD-L	CVM	CVM-L	KS	KS-L	LRT	KL
IG(1)	30	0.36	0.25	0.32	0.37	0.29	0.38	0.32	**0.38**
	50	0.51	0.42	0.41	0.53	0.39	0.54	0.53	**0.58**
	100	0.83	0.81	0.81	0.84	0.66	0.85	0.87	**0.89**
IG(1.5)	30	0.20	0.23	0.19	0.24	0.19	0.23	0.18	**0.26**
	50	0.30	0.33	0.27	0.37	0.21	0.38	0.28	**0.40**
	100	0.68	0.69	0.65	0.72	0.48	0.73	0.74	**0.77**
LN(0.8)	30	0.21	0.21	0.17	**0.31**	0.10	0.31	0.21	0.30
	50	0.30	0.33	0.30	0.39	0.25	0.41	0.31	**0.42**
	100	0.60	0.60	0.57	0.60	0.47	0.64	0.61	**0.68**
GO(2)	30	0.32	**0.54**	0.31	0.43	0.28	0.42	0.30	0.30
	50	0.59	**0.79**	0.57	0.75	0.48	0.78	0.64	0.69
	100	0.91	**0.99**	0.88	**0.99**	0.70	**0.99**	0.95	0.98
GO(4)	30	0.49	**0.66**	0.45	0.53	0.39	0.52	0.47	0.41
	50	0.70	**0.85**	0.68	0.81	0.49	0.80	0.75	0.74
	100	0.96	**1.00**	0.96	**1.00**	0.91	**1.00**	0.98	0.99
LF(2)	30	0.15	**0.24**	0.15	0.20	0.15	0.17	0.21	0.14
	50	0.24	**0.43**	0.20	0.29	0.18	0.32	0.26	0.23
	100	0.42	**0.60**	0.40	0.58	0.39	0.58	0.50	0.50
LF(4)	30	0.19	**0.32**	0.18	0.16	0.16	0.23	0.18	0.13
	50	0.16	**0.35**	0.16	0.24	0.11	0.26	0.16	0.14
	100	0.58	**0.81**	0.51	0.74	0.41	0.80	0.63	0.65

Table 3: Powers of eight test methods for the Gamma distribution under different alternatives.

p	Model	n=100			n=200			n=300		
		Bias	Var	KL	Bias	Var	KL	Bias	Var	KL
1	Cubic	40.51	2.32	42.84	50.69	1.58	52.27	43.86	1.11	44.97
	L-spline	7.51	1.67	9.18	4.72	0.88	5.60	1.85	0.54	2.39
2	Cubic	44.39	1.72	46.11	44.55	1.01	45.55	45.91	0.82	46.72
	L-spline	13.59	1.49	15.07	9.07	0.89	9.96	4.07	0.82	4.89
3	Cubic	46.11	1.44	47.55	51.23	0.85	52.08	54.30	0.70	55.00
	L-spline	65.28	1.86	67.15	29.34	1.79	31.13	35.74	0.45	36.19

Table 4: Biases, variances, and KL distances in 10^{-2} with the GIGG family.

with $\zeta = -0.5$. We set $\alpha_0 = 3$ and $\alpha_1 = 3$ in the simulation, and consider five choices of ζ: $\zeta = -3, -2, -0.5, 2$ and 3 which reflect different degrees of departure from the IG distribution. Table 5 lists powers of seven test methods with significance level set at 5%. The AD-L statistic cannot be calculated since the estimate of $F_0(x)\,(1 - F_0(x))$ is close to zero. The powers are the type I errors when $\zeta = -0.5$. It is clear that all methods have type I error smaller or close to 5%. Again, the modified CVM and KS tests have larger powers than those from the original tests.

7 Conclusion

In this paper, we proposed model-based penalities for smoothing spline density estimation and inference. It successfully incorporates indefinite prior information about the density in density estimation and inference process. Two examples, respectively from Pearson and GGIG family, are used to show the derivation. The simulation results in Table 1 and 4 show the reduction of KL divergence, including both bias and variance, of density estimation from the new model-based penalities. And Table 2, 3 and 5 show the modification of test power using the proposed model.

ζ	Sample Size	AD	CVM	CVM-L	KS	KS-L	LRT	KL
	300	0.05	0.03	0.06	0.06	0.06	0.03	0.08
-0.5	200	0.06	0.05	0.01	0.05	0.02	0.04	0.02
	100	0.04	0.04	0.02	0.03	0.03	0.03	0.05
	300	0.54	0.48	**0.67**	0.36	**0.67**	0.54	0.54
3	200	0.37	0.32	**0.44**	0.29	**0.42**	0.35	0.35
	100	0.19	0.2	**0.24**	0.1	0.22	0.15	0.16
	300	0.34	0.32	**0.37**	0.22	0.36	0.26	0.28
2	200	0.34	0.29	**0.38**	0.25	**0.38**	0.25	0.29
	100	0.16	0.15	0.18	0.11	**0.19**	0.12	0.08
	300	0.19	0.21	**0.34**	0.18	**0.34**	0.21	0.21
-3	200	0.16	0.17	**0.23**	0.15	**0.23**	0.17	0.19
	100	0.14	0.12	0.14	0.12	0.14	**0.17**	0.15
	300	0.06	0.06	**0.14**	0.06	**0.14**	0.06	0.06
-2	200	0.13	0.14	**0.15**	0.13	0.14	0.08	0.08
	100	0.12	0.12	0.11	0.08	**0.12**	0.09	0.09

Table 5: Powers of seven test methods for the IG distribution.

References

Anderson, T. W. & Darling, D. A. (1954). A test of goodness of fit, *Journal of the American Statistical Association* **49**(268): 765–769.

Del Castillo, J. & Puig, P. (1997). Testing departures from gamma, rayleigh and truncated normal distributions, *Annals of the Institute of Statistical Mathematics* **49**(2): 255–269.

Gu, C. (2013). *Smoothing Spline ANOVA Models, 2nd ed.*, Springer-Verlag, New York.

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

Gu, C. & Wang, J. (2003). Penalized likelihood density estimation: Direct cross validation and scalable approximation, *Statistica Sinica* **13**: 811–826.

Heckman, N. & Ramsay, J. O. (2000). Penalized regression with model-based penalties, *Canadian Journal of Statistics* **28**: 241–258.

Heskes, T. (1998). Bias/variance decompositions for likelihood-based estimators, *Neural Computation* **10**(6): 1425–1433.

Lehmann, E. L. & Romano, J. P. (2006). *Testing Statistical Hypotheses*, Springer Science & Business Media.

Loader, C. R. (1999). *Local Regression and Likelihood*, Springer, New York.

Pearson, K. (1894). Contributions to the mathematical theory of evolution, *Philosophical Transactions of the Royal Society of London. A* **185**: 71–110.

Romantsova, Y. V. (1996). On an asymptotic goodness-of-fit test for a two-parameter gamma-distribution, *Journal of Mathematical Sciences* **81**(4): 2759–2765.

Shakil, M., Kibria, B. G. & Singh, J. N. (2016). A new family of distributions based on the generalized pearson differential equation with some applications, *Austrian Journal of Statistics* **39**(3): 259–278.

Silverman, B. W. (1984). Spline smoothing: the equivalent variable kernel method, *Annals of Statistics* **12**: 898–916.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, New York.

Stephens, M. A. (1974). EDF statistics for goodness of fit and some comparisons, *Journal of the American Statistical Association* **69**(347): 730–737.

Stephens, M. A. (1986). *Tests Based on EDF Statistics*, Vol. 68 of *Statistics: Textbooks and Monographs*, Marcel Dekker, Inc., New York, chapter 4, pp. 97–191.

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

Wang, Y. (2011). *Smoothing Splines: Methods and Applications*, Chapman and Hall, New York.

Appendix A Derivation of a Reproducing Kernel for the Gamma Distribution

To save space we show a brief derivation of the RK for the Gamma distribution. Since $\mathcal{H}_0 = \{1, x, \log(x)\}$ (the constant function will be removed after this construction), the Wronskian matrix is

$$W(x) = \begin{bmatrix} 1 & x & \log(x) \\ 0 & 1 & 1/x \\ 0 & 0 & -\frac{1}{x^2} \end{bmatrix}, \tag{A.1}$$

and

$$W^{-1}(x) = \begin{bmatrix} 1 & -x & -x^2 + x^2\log(x) \\ 0 & 1 & x \\ 0 & 0 & -x^2 \end{bmatrix}. \tag{A.2}$$

The Green function is

$$G(t,s) = -s^2 + s^2\log(s) + ts - s^2\log(t) \quad (s <= t). \tag{A.3}$$

Thus, the RK of $\mathcal{H}_1$ is

$$\begin{aligned} R_1(x,z) &= \int_0^T G(x,s)G(z,s)ds \\ &= (1 + \log(z) + \log(x) + \log(z)\log(x)) * I_4(x \wedge z) \\ &\quad - (z + x + z\log(x) + x\log(z)) * I_3(x \wedge z) \\ &\quad + xzI_2(x \wedge z) \\ &\quad + I_{4,2}(x \wedge z) \\ &\quad - (2 + \log(z) + \log(x))\, I_{4,1}(x \wedge z) \\ &\quad + (z + x)I_{3,1}(x \wedge z), \end{aligned} \tag{A.4}$$

where $x \wedge z = \min(x, z)$, and

$$\begin{aligned} I_p(s) &= \int_0^s x^p dx = \frac{1}{p+1}(s)^{p+1} \\ I_{p,k}(s) &= \int_0^s x^p \log(x)^k = \frac{1}{p+1}(s)^{p+1}\log(s)^k - \frac{k}{p+1}I_{p+1,k-1}(s). \end{aligned} \tag{A.5}$$

Appendix B Derivation of a Reproducing Kernel for the Beta Distribution

To save space we show a brief derivation of the RK for the Beta distribution. Given the differential operator L in equation (11), the Wronskian matrix asscociated with $\mathcal{H}_0$ is

$$W(x) = \begin{bmatrix} 1 & \log(x) & \log(1-x) \\ 0 & 1/x & 1/(x-1) \\ 0 & -\frac{1}{x^2} & -\frac{1}{(x-1)^2} \end{bmatrix}, \tag{A.6}$$

and

$$W^{-1}(x) = \begin{bmatrix} 1 & (x-1)^2\log(1-x) - x^2\log(x) & x(x-1)\left[(x-1)\log(1-x) - x\log(x)\right] \\ 0 & x^2 & x^2(x-1) \\ 0 & -(x-1)^2 & -x(x-1)^2 \end{bmatrix}. \quad \text{(A.7)}$$

The Green function is

$$G(t,s) = s(s-1)\left[(s-1)\log(1-s) - s\log(s)\right] + s^2(s-1)\log(t) - s(s-1)^2\log(1-t), \quad s <= t. \quad \text{(A.8)}$$

Then, the RK of $\mathcal{H}_1$ is

$$\begin{aligned} R_1(x,z) &= \int_0^{x\wedge z} G(x,s)G(z,s)ds \\ &= (\log(z)\log(1-x) + \log(x)\log(1-z))I(x\wedge z;3,3,0,0) \\ &\quad + \log(1-x)\log(1-z)I(x\wedge z;2,4,0,0) + \log(x)\log(z)I(x\wedge z;4,2,0,0) \\ &\quad + I(x\wedge z;2,4,0,2) + (\log(x)+\log(z))I(x\wedge z;3,3,0,1) \\ &\quad - (\log(1-x)+\log(1-z))I(x\wedge z;2,4,0,1) \\ &\quad + I(x\wedge z;4,2,2,0) - (\log(x)+\log(z))I(x\wedge z;4,2,1,0) \\ &\quad - (\log(1-x)+\log(1-z))I(x\wedge z;3,3,1,0) \\ &\quad + 2I(x\wedge z;3,3,1,1), \end{aligned} \quad \text{(A.9)}$$

where

$$I(y;m_1,m_2,m_3,m_4) = \int_0^y x^{m_1}(1-x)^{m_2}\log(x)^{m_3}\log(1-x)^{m_4}dx. \quad \text{(A.10)}$$

Appendix C Derivation of a Reproducing Kernel for the GGIG Family

To save space we show a brief derivation of the RK for the GGIG family with $p = 1$ only. Note that $L = D^4 + 6x^{-1}D^3 + 6x^{-2}D^2$ and $\mathcal{H}_0 = \text{span}\{1, \log(x), x, x^{-1}\}$. The Wronskian matrix asscociated

with $\mathcal{H}_0$ is

$$W(x) = \begin{bmatrix} 1 & x & x^{-1} & \log(x) \\ 0 & 1 & -x^{-2} & x^{-1} \\ 0 & 0 & 2x^{-3} & -x^{-2} \\ 0 & 0 & -6x^{-4} & 2x^{-3} \end{bmatrix}, \tag{A.11}$$

and

$$W^{-1}(x) = \begin{bmatrix} 1 & -x & -x^2 + 3x^2\log(x) & x^3\log(x) \\ 0 & 1 & 2x & .5x^2 \\ 0 & 0 & -x^3 & -.5x^4 \\ 0 & 0 & -3x^2 & -x^3 \end{bmatrix}. \tag{A.12}$$

The Green function is

$$G(t,s) = -\frac{s^4}{2t} + \frac{s^2 t}{2} + s^3\log(s) - s^3\log(t), \quad s \le t. \tag{A.13}$$

Thus, the RK of $\mathcal{H}_1$

$$\begin{aligned} R_1(x,z) &= \int_0^T G(x,s)G(z,s)ds \\ &= \frac{1}{36xz}(x\wedge z)^9 - \frac{1}{16}(x\wedge z)^8\log(x\wedge z)\left(\frac{1}{x}+\frac{1}{z}\right) \\ &\quad + \frac{1}{16}(x\wedge z)^8\left(\frac{1}{8x}+\frac{1}{8z}+\frac{1}{z}\log(x)+\frac{1}{x}\log(z)\right) \\ &\quad - \frac{1}{7}(x\wedge z)^7\log(x\wedge z)\left(\frac{2}{7}+\log(x)+\log(z)\right) + \frac{1}{7}(x\wedge z)^7\log(x\wedge z)^2 \\ &\quad + \frac{1}{7}(x\wedge z)^7\left(\frac{2}{49}-\frac{z}{4x}-\frac{x}{4z}+\frac{1}{7}\log(x)+\frac{1}{7}\log(z)+\log(x)\log(z)\right) \\ &\quad + \frac{1}{12}(x+z)(x\wedge z)^6\log(x\wedge z) - \frac{1}{12}(\frac{x}{6}+x\log(z)+\frac{z}{6}+z\log(x))(x\wedge z)^6 + \frac{1}{20}(x\wedge z)^5 xz. \end{aligned}$$