WHAT DO KERNEL DENSITY ESTIMATORS OPTIMIZE?

ROGER KOENKER, IVAN MIZERA, AND JUNGMO YOON

Abstract. Some linkages between kernel and penalty methods of density estimation are explored. It is recalled that classical Gaussian kernel density estimation can be viewed as the solution of the heat equation with initial condition given by data. We then observe that there is a direct relationship between the kernel method and a particular penalty method of density estimation. For this penalty method, solutions can be characterized as a weighted average of Gaussian kernel density estimates, the average taken with respect to the bandwidth parameter. A Laplace transform argument shows that this weighted average of Gaussian kernel estimates is equivalent to a fixed bandwidth kernel estimate using a Laplace kernel. Extensions to higher order kernels are considered and some connections to penalized likelihood density estimators are made in the concluding sections.

1. Introduction

In economics it is commonly believed that a phenomenon is understood if and only if one can formulate an optimization problem whose solution emulates the phenomenon. In this respect the appeal and apparent success of kernel density estimation methods in statistics, and especially in econometrics, is something of an anomaly. By partially answering the question posed in the title we hope to shed some light on the rationale underlying kernel estimation. We make no great claims for the novelty of our account, indeed many aspects will be familiar to those conversant with the regularization literature, especially the influential papers of Silverman (1982, 1984a), and the somewhat less easily accessible papers of Terrell (1990) and Aidu and Vapnik (1989), but these ideas may be less familiar within the broader statistical community.

The linearization, or equivalent kernel characterization, of L_2 penalty methods for nonparametric regression has proven to be a valuable device for studying their asymptotic behavior, as vividly demonstrated by Silverman (1984b), and the recent paper of Li and Ruppert (2008). Our objective is more modest: reexamining kernel methods from the optimization perspective exposes some peculiarities of the kernel approach and suggests some attractive alternatives within the penalization framework.

Version: July 5, 2009. Roger Koenker is McKinley Professor of Economics and Professor of Statistics, University of Illinois, Urbana, IL 61801 USA. Ivan Mizera is Professor of Mathematical and Statistical Sciences, University of Alberta, Edmonton, Alberta, T6G 2G1 Canada. Jungmo Yoon is Assistant Professor of Economics, Claremont-McKenna College, Claremont, CA 91711, USA. This research was partially supported by NSF grant SES-05-44673, and by the Natural Sciences and Engineering Research Council of Canada.

2. Gaussian Kernel Density Estimation via the Heat Equation

Consider the proverbial rod of infinite length. Denote temperature of the rod at a point x and time t by $\phi(x,t)$. Let $\phi_x(x,t), \phi_{xx}(x,t)$ be the first and second derivatives with respect to its first argument, and $\phi_t(x,t)$ be the first derivative with respect to the second argument. If the initial temperature is described by a function $g(x)$, the temperature at (x,t) is determined by the heat equation,

$$\phi_t(x,t) = \phi_{xx}(x,t) \quad -\infty < x < \infty, \;\; t > 0, \tag{2.1}$$

with initial condition,

$$\phi(x,0) = g(x) \quad -\infty < x < \infty,$$

and boundary conditions,

$$\phi(x,t) \to 0 \quad \text{as} \;\; x \to \pm\infty.$$

The solution of the heat equation is, see e.g. Strauss (1992),

$$\phi(x,t) = \frac{1}{\sqrt{4\pi t}} \int_{-\infty}^{\infty} g(z) \exp(-\frac{(x-z)^2}{4t}) dz.$$

Now suppose the initial condition $g(x)$ is given by a sum of (Dirac) point masses, $g(x) = n^{-1} \sum_{i=1}^n \delta_{X_i}(x)$, then the solution takes the form

$$\phi(x,t) = \frac{1}{\sqrt{4\pi t}} \int_{-\infty}^{\infty} \frac{1}{n} \sum_{i=1}^n \delta_{X_i}(z) \exp(-\frac{(x-z)^2}{4t}) dz = \frac{1}{\sqrt{4\pi t} n} \sum_{i=1}^n \exp(-\frac{(x-X_i)^2}{4t})$$

This solution is immediately recognizable as a Gaussian kernel density estimate with the value $\sqrt{2t}$ playing the role of the bandwidth. As time passes and the heat diffuses through the rod, its distribution at time t is precisely given by a Gaussian kernel density estimate. This diffusion interpretation of kernel smoothing is conventional in the imaging literature where partial differential equation methods are commonplace, but it is perhaps less familiar elsewhere. In statistics, the multi-resolution work of Chaudhuri and Marron (2000) constitutes an important exception.

3. A Roughness Penalty Interpretation of the Kernel Method

Consider the integral transform of $\phi(x,t)$

$$f(x,\lambda) = \int_0^{\infty} \tfrac{1}{\lambda} e^{-t/\lambda} \phi(x,t) dt.$$

for a fixed $\lambda > 0$. We will show that the variational problem,

$$\min_f \int [\tfrac{\lambda}{2} f_x(x,\lambda)^2 + \tfrac{1}{2} f(x,\lambda)^2 - \phi(x,0) f(x,\lambda)] dx \tag{3.1}$$

leads back to the solution of heat equation and therefore to the kernel estimate. To see this, consider the integral transform of both sides of heat equation (2.1)

$$\int_0^\infty \tfrac{1}{\lambda} e^{-t/\lambda} \phi_t(x,t)dt = \int_0^\infty \tfrac{1}{\lambda} e^{-t/\lambda} \phi_{xx}(x,t)dt.$$

Integrating by parts,

$$\tfrac{1}{\lambda} e^{-t/\lambda} \phi(x,t)|_0^\infty + \tfrac{1}{\lambda} \int_0^\infty \tfrac{1}{\lambda} e^{-t/\lambda} \phi(x,t)dt = \frac{\partial^2}{\partial x^2} \int_0^\infty \tfrac{1}{\lambda} e^{-t/\lambda} \phi(x,t)dt$$

and we have,

$$-\phi(x,0) + f(x,\lambda) = \lambda f_{xx}(x,\lambda) \tag{3.2}$$

By applying the integral transform on the boundary condition of the heat equation, we obtain a boundary condition $f(x,\lambda) \to 0 \quad \text{as} \quad x \to \pm\infty$. Finally, note that the variational integral (3.1) can be minimized by solving the Euler equation,

$$\frac{\partial \Psi}{\partial f} - \frac{\partial}{\partial x}\frac{\partial \Psi}{\partial f_x} = 0 \quad \text{where} \quad \Psi = \tfrac{\lambda}{2} f_x^2 + \tfrac{1}{2} f^2 - \phi(x,0) f.$$

Since $\partial\Psi/\partial f = f - \phi(x,0)$ and $\partial\Psi/\partial f_x = \lambda f_x$, we obtain equation (3.2). Because (3.2) is equivalent to the heat equation representation (2.1), we see that the solution of the minimization problem (3.1) is the integral transform of the solution of heat equation (2.1). See *e.g.*, Carrier and Pearson (1988), for some further details of this argument.

Note that $f(x,\lambda)$ itself is a proper density function since

$$\int_{-\infty}^\infty f(x,\lambda)dx = \int_{-\infty}^\infty \int_0^\infty \tfrac{1}{\lambda} e^{-t/\lambda} \phi(x,t)dtdx = \int_0^\infty \tfrac{1}{\lambda} e^{-t/\lambda} dt \cdot \int_{-\infty}^\infty \phi(x,t)dx = 1.$$

As before, let data describe the initial condition so, $\phi(x,0) = g(x) = n^{-1}\sum_{i=1}^n \delta_{X_i}(x)$ in (3.1). With $f(x,\lambda)$ as our density estimate, we can write (3.1) as,

$$\min_f - \int f(x,\lambda) dF_n(x) + \tfrac{1}{2} \int (f(x,\lambda))^2 dx + \tfrac{\lambda}{2} \int (f_x(x,\lambda))^2 dx. \tag{3.3}$$

The first two terms can be interpreted as a measure of infidelity to the observed data, the last term is a penalty on the roughness of the fitted density. The parameter λ plays the role of a tuning, or regularization, parameter. Our interpretation of each of these components probably warrants some additional explanation.

To interpret the fidelity term, first note that if we were to replace dF_n by a limiting form of the true density, then minimizing just the fidelity term with respect to f would have to reproduce this density. Second, consider a discretization. Instead of integrating with respect to $dF_n(x)$ suppose that we approximate the empirical measure by a piecewise constant density with respect to Lebesgue measure, say $f_n(x)$. We can write the mean

squared error criterion,

$$\int (f(x) - f_n(x))^2 dx = \int f^2 dx - 2\int f f_n dx + \int f_n^2 dx.$$

Since we are minimizing with respect to f, the last term can be neglected. The connection to the Pearson minimum χ^2 framework is more apparent if we write,

$$E_f[\frac{(f-f_n)^2}{f}] = \int \frac{(f-f_n)^2}{f} f dx = \int (f-f_n)^2 dx.$$

So we conclude that the fidelity term implicit in Gaussian kernel density estimators is a continuous analogue of Pearson's minimum χ^2 approach. Just as the discrete formulation of the Pearson criterion attempts to minimize the difference between the observed and the expected frequencies, penalized density estimation method try to minimize distance between a density estimate and the empirical density, distance measured by the least squares principle. This approach is closely related to the "histospline" approach of Boneva, Kendall, and Stefanov (1971).

The roughness penalty appearing in (3.3) is also somewhat strange. More typically roughness penalties would be based on curvature, that is on second derivatives of the density. The early penalized density estimator of Good and Gaskins (1971) was based on the first derivative, but it was the first derivative of the square root of the density, a quantity that when squared and then integrated yields Fisher information. In the present case the penalty is somewhat simpler and does not appear to have an obvious statistical or probabilistic interpretation.

Regarding the tuning parameter λ, observe that the solution of the penalty method (3.3) is an integral transform of a family of density function estimates $\{\phi(x,t)\}_{t\in[0,\infty)}$ indexed by the smoothing parameter t. Solutions can be interpreted as weighted averages of various kernel density estimates for various bandwidth t, with the weight determined by an exponential distribution with intensity parameter λ. The tuning parameter λ thus determines the relative weight of kernel density estimates $\{\phi(x,t)\}_{t\in[0,\infty)}$ when we calculate the weighted average. Small λ, puts more weight on the kernel density estimates with smaller bandwidths t, so we obtain a rather rough density, while larger λ's yield a smoother density.

Thus far we have presented a penalized density estimation problem corresponding to the Gaussian kernel density estimator. But we have not exhibited an explicit solution of the penalty problem. Although it is usually true of penalty methods that solutions are defined only implicitly as a solutions of a variational problem, and thus need to be computed by some iterative algorithm, in the present instance we can present an explicit form of the solution. Solving (3.1) is equivalent to solving equation (3.2). We can rewrite (3.2) as,

$$-\lambda f_{xx}(x,\lambda) + f(x,\lambda) = dF_n(x). \tag{3.4}$$

To solve (3.4), define the Green's function which satisfies

$$-\lambda G_{xx}(x) + G(x) = \delta_0(x) \tag{3.5}$$

with a boundary condition $G(x) \to 0$ as $x \to \pm\infty$. One can show that either by direct integration or using the Fourier transform, that the equation (3.5) together with its boundary condition has the following solution,

$$G(x) = \frac{1}{2\sqrt{\lambda}} \exp(-|x|/\sqrt{\lambda}).$$

The solution of (3.4) is then obtained by convoluting the Green's function with the right hand side of the equation (3.4),

$$\hat{f}(x,\lambda) = \int G(x-z)dF_n(z) = \frac{1}{2\sqrt{\lambda}n} \sum_{i=1}^{n} \exp(-|x-X_i|/\sqrt{\lambda}). \tag{3.6}$$

We have thus obtained the following explicit kernel representation of the penalized density estimator.

Theorem 1. *The solution of* (3.3) *is given by*

$$\hat{f}_\lambda(x) = \frac{1}{2\sqrt{\lambda}n} \sum_{i=1}^{n} \exp(-|x-X_i|/\sqrt{\lambda}), \tag{3.7}$$

a kernel density estimate with the double-exponential (Laplacian) kernel and bandwidth $\sqrt{\lambda}$.

Remark 1: A natural question would be whether this process is reversible, that is, whether one can recover Gaussian kernel density estimate from the solution of the penalty method. In order to investigate this it is convenient to consider the unnormalized density $\lambda f(x,\lambda) = \int_0^\infty e^{-t/\lambda}\phi(x,t)dt$ with mass λ. We may regard $\lambda f(x,\lambda)$ as the Laplace transform of a family of Gaussian kernel density estimates $\phi(x,t)$. In doing so, we can exploit well-established relationships of Laplace transform and their inverse transforms. Note, however, that this Laplace transform is defined in terms of the bandwidth parameter t *not* x. From Theorem 1, we know

$$\lambda\hat{f}(x,\lambda) = \frac{\sqrt{\lambda}}{2n} \sum_{i=1}^{n} \exp(-|x-X_i|/\sqrt{\lambda}) \tag{3.8}$$

Now apply the inverse Laplace transform to (3.8). Since the inverse Laplace transform of the double exponential function is Gaussian, applying the inverse Laplace transform to (3.8) term-by-term, we obtain the original kernel density estimate

$$\phi(x,t) = \frac{1}{n\sqrt{4\pi t}} \sum_{i=1}^{n} \exp(-(x-X_i)^2/4t).$$

So we can indeed recover the original Gaussian kernel density function estimate. ■

Remark 2: A final interpretation can be seen by considering the family of normal mixtures,

$$f(x) = \int\int (\sqrt{4\pi t})^{-1} \exp(-(x-z)^2/4t) dQ_1(z) dQ_2(t)$$

where $dQ_1(\cdot)$ is the mixing distribution of the location parameters and $dQ_2(\cdot)$ is the mixing distribution of the scale parameter. If we take $dQ_1(z) = dF_n(z)$ and $dQ_2(t) = \lambda^{-1} e^{-t/\lambda} dt$, then the normal mixture density representation yields $\hat{f}_\lambda(x)$. Since the mixing distribution for location parameters results in the Gaussian kernel density estimate, the penalized density estimator (3.7) can be interpreted as a scale mixture of the Gaussian kernel density estimates. ■

Remark 3: Other kernel estimators can be derived from modifications of the partial differential equation formulation of the variational solution. For instance, the Epanechnikov kernel, $K(x) = \frac{3}{4}(1-x^2)I(|x| \leq 1)$ can be shown to arise from the nonlinear diffusion equation,

$$\phi_t = (\phi\phi_x)_x \quad -\infty < x < \infty, \;\; t > 0, \tag{3.9}$$

subject to initial conditions $\phi(x,0) = \delta_0(x) \quad -\infty < x < \infty$, boundary conditions, $\phi(x,t) \to 0 \quad \text{as} \quad x \to \pm\infty$, and the condition that $\int \phi(x,t)dx = 1$ for all $t > 0$. See, for example, Exercise 5.3.6 of McOwen (2003). ■

Starting from the heat equation whose solution is the Gaussian kernel density estimate, we have exhibited a variational problem that shares some common structure with the heat equation. We argued that the variational problem can be interpreted as the penalized minimum χ^2 method of density estimation. The solution of the penalty problem can be characterized in terms of the Gaussian kernel method, as a weighted average of Gaussian kernel estimates with exponentially declining weights with respect to bandwidth. In addition, this weighted average of Gaussian kernel density estimates can be represented as a conventional fixed bandwidth kernel density estimator employing a Laplace (double exponential) kernel. This leads one to ask: how reasonable is the fidelity and roughness penalty that define the penalty problem? Neither, one would have to say, is very appealing; we will briefly explore some alternatives in the final two sections of the paper.

4. Higher-Order Kernels and Derivative Penalties

Thus far we have seen that the $\mathcal{L}_2$ roughness penalty,

$$P(f) = \int (f'(x))^2 dx,$$

when combined with a simple Pearsonian measure of fidelity yields solutions that can be interpreted as classical kernel density estimators. Since it is evident more generally that such quadratic variational problems have solutions represented by linear operators that

must also have kernel interpretations it seems obligatory to press ahead with the question: What happens with higher-order derivative penalties?

Consider the second-order penalty problem

$$\min_f \{\tfrac{1}{2}\int f^2(x)dx - \int f(x)dF_n(x) + \tfrac{\lambda}{2}\int (f''(x))^2 dx\}. \tag{4.1}$$

As in the previous case we can express the solution of this problem as a kernel estimator with fixed bandwidth depending on λ, except that we now require a "higher-order" kernel.

Theorem 2. *Solutions* (4.1) *have the kernel representation,*

$$\hat{f}_\sigma(x) = \int G_\sigma(x-z)dF_n(z).$$

where $G_\sigma(u) = \frac{1}{2}\exp(-|u/\sigma|/\sqrt{2})\sin(|u/\sigma|/\sqrt{2}+\pi/4)/\sigma$, *and* $\sigma = \lambda^{1/4}$.

Proof: The Euler condition for the problem (4.1) is,

$$\lambda f_\lambda^{(iv)}(x) + f_\lambda(x) = dF_n(x)$$

It is easily verified that the Green's function associated with this differential equation satisfying,

$$\lambda G^{(iv)}(x) + G(x) = \delta_0(x)$$

takes the asserted form, with boundary conditions $G(x) \to 0$ and $G'(x) \to 0$ as $x \to \pm\infty$ and consequently solutions have the integral representation appearing in the theorem. ∎

This result is a particularly simple example of the general theory of reproducing kernel Hilbert spaces as expounded for example by Wahba (1990). The function G is a second-order kernel satisfying the condition, $\mu_k = \int u^k G(u) = 0$ for $i = 1, 2, 3$, and $\mu_0 = 1$; it is precisely the kernel derived by Silverman (1984b) to approximate the penalized likelihood estimator with conventional roughness penalty:

$$P(f) = \int (\log f''(x))^2 dx,$$

except that we are now penalizing roughness of the density itself, rather than the roughness of the logarithm of the density. In contrast to Silverman's setting where this kernel provides an approximation to the penalized maximum likelihood estimator, here there is an exact equivalence.

The result given in Theorem 2 can be extended to yet higher order derivative penalties yielding yet higher order kernels. In Figure 1 we illustrate the kernels appearing in Theorems 1 and 2 as well as the kernel corresponding to the third order derivative penalty proposed by Silverman (1982).

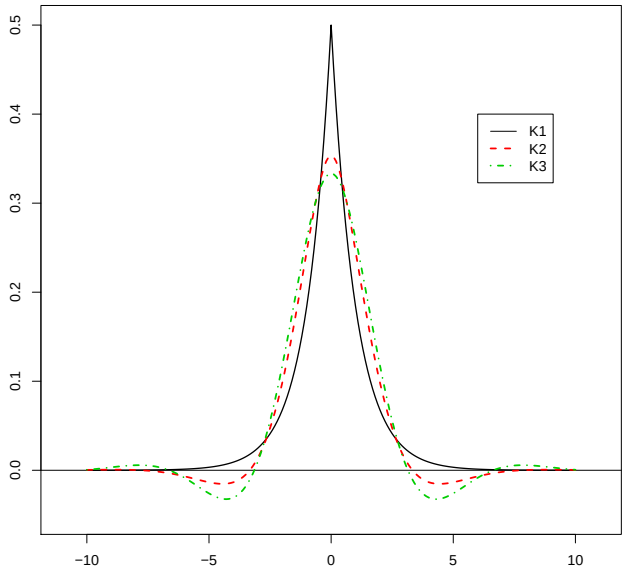

FIGURE 1. Equivalent Kernels for L_2 Density Estimation Penalties

We can also ask whether there is an analogue of the diffusion representation of this second-order penalty estimator. Consider the modified diffusion equation,

$$\phi_t(x,t) = -\phi_{xxxx}(x,t) \quad -\infty < x < \infty, \quad t > 0, \tag{4.2}$$

with initial condition, $\phi(x,0) = g(x) \quad -\infty < x < \infty$, and boundary conditions, $\phi(x,t) \to 0$ and $\phi'(x,t) \to 0$ as $x \to \pm\infty$. Let

$$\Phi(\xi,t) = \int e^{-2\pi i\xi x}\phi(x,t)dx$$

denote the Fourier transform of ϕ. Integrating by parts repeatedly we can rewrite the Fourier transform of (4.2),

$$\int e^{-2\pi i\xi x}\phi_t(x,t)dx = -\int e^{-2\pi i\xi x}\phi_{xxxx}(x,t)dx$$

as

$$\frac{\partial}{\partial t}\Phi(\xi,t) = -16\pi^4\xi^4 t\ \Phi(\xi,t).$$

But this elementary first order differential equation has solution,

$$\Phi(\xi,t) = \Phi(\xi,0)e^{-16\pi^4\xi^4 t}$$

where $\Phi(\xi,0) = \int \phi(x,0)e^{-2\pi i\xi x}dx$. Thus, inverting the Fourier transform, we obtain the convolution,

$$\phi(x,t) = \phi(x,0) * K(x,t) = \int_{-\infty}^{\infty} g(z)K(x-z,t)dz$$

where $K(x,t) = K(x/t,1)/t$ and

$$\begin{aligned} K(x,1) &= \int e^{2\pi i\xi x}e^{-16\pi^4\xi^4}d\xi \\ &= \frac{1}{2\pi}\left(2\Gamma(\tfrac{5}{4})\ {}_0H_2[\{\},\{\tfrac{1}{2},\tfrac{3}{4}\},\tfrac{x^4}{256}] - \tfrac{1}{4}x^2\Gamma(\tfrac{3}{4})\ {}_0H_2[\{\},\{\tfrac{5}{4},\tfrac{3}{2}\},\tfrac{x^4}{256}]\right) \end{aligned}$$

and ${}_pH_q\left[\{a_1,\cdots a_p\},\{b_1,\cdots b_q\},x\right]$ is the generalized hypergeometric function. The kernel is again a second-order kernel. As in the previous section we can interpret the function $\phi(x,t)$ as describing the state of the diffusion at time t. Integrating out t by averaging over these solutions with exponentially declining weights yields the solution given in Theorem 2.

5. Penalized Likelihood Methods

In contrast to the foregoing approach in which fidelity is represented by a Pearsonian squared error criterion, the tradition originating in Good and Gaskins (1971), and further developed by de Montricher, Tapia, and Thompson (1975), Silverman (1982), Cox and O'Sullivan (1990), and Eggermont and LaRiccia (1999) is to pose the density estimation problem as a maximum likelihood problem subject to a penalty on the roughness of the fitted density.

Eggermont and LaRiccia (1999) demonstrate that one can approximate the penalized likelihood estimator of Good and Gaskins (1971), by a kernel estimator, and then use the known properties of kernel estimators to investigate asymptotic behavior of the penalty method. We briefly describe this approach and its connections to the results of the previous section. The Good and Gaskins (1971) penalized likelihood estimator solves,

$$\min_v\{-2\int \log v(x)dF_n(x) + \int (v(x))^2dx + \lambda\int (v_x(x))^2dx\}, \tag{5.1}$$

where, v can be interpreted as the square root of the estimated density, and the negative log-likelihood function is employed as a measure of fidelity. The second term looks similar to the penalty term of (3.3), but it plays a completely different role, of enforcing a normalization constraint. The third term is the roughness penalty. Given the interpretation of v the penalty is proportional to the Fisher information for the location parameter of the density. One may well ask why Fisher information for location is a reasonable measure of roughness of densities, but we will not attempt to defend this choice here.

Invoking the Euler equation, de Montricher, Tapia, and Thompson (1975) showed that the solutions of (5.1) satisfy the second order non-linear differential equation

$$-\lambda v_{xx}(x) + v(x) = \frac{dF_n(x)}{v(x)} \tag{5.2}$$

with boundary condition $v(x) \to 0$ as $x \to \pm\infty$. To obtain a solution of (5.2), we can apply the same principle that we used in (3.4). The Green's function is

$$G_{\sqrt{\lambda}}(x) = \tfrac{1}{2\sqrt{\lambda}} \exp(\tfrac{-|x|}{\sqrt{\lambda}})$$

and convolution of the right hand side of (5.2) and the Green's function yields,

$$\hat{v}(x) = \int G_{\sqrt{\lambda}}(x-z)\frac{dF_n(z)}{v(z)} = \frac{1}{n}\sum_{i=1}^{n} \tfrac{1}{2\sqrt{\lambda}} \exp(\tfrac{-|x-X_i|}{\sqrt{\lambda}})/\hat{v}(X_i).$$

Note that this is *not* a solution of the differential equation (5.2), since the right hand side is a function of the unknown $\hat{v}$. We have simply converted a differential equation into an integral equation. Finding an analytic solution of (5.2) is difficult because of this nonlinearity. Rather than solving the equation directly, Eggermont and LaRiccia (1999) devised an elegant way to construct upper and lower bounds for density estimator arising from equation (5.2). We will summarize their methods briefly.

Rewriting (5.2), using $(v^2)_{xx} = 2vv_{xx} + 2(v_x)^2$ we have,

$$\begin{aligned} -\tfrac{\lambda}{2}(v^2)_{xx} + v^2 &= dF_n - \lambda(v_x)^2 \\ -\tfrac{\lambda}{4}(v^2)_{xx} + v^2 &= \tfrac{1}{2}dF_n + \tfrac{1}{2}(1 - \lambda(v_x/v)^2)(v_x)^2 \end{aligned}$$

Bearing in mind that it was $\hat{v}^2$ from solving (5.1) that was intended to be a density, we can set $\hat{f}(x) = \hat{v}^2(x)$ and employ the Green's functions for each case. Then ignoring the second components of the right hand side of each equation, one obtains upper and lower bounds for the density estimate:

$$\tfrac{1}{2}\int G_{\sqrt{\lambda/4}}(x-z)dF_n(z) \le \hat{f}(x) \le \int G_{\sqrt{\lambda/2}}(x-z)dF_n(z),$$

or more explicitly,

$$\frac{1}{2n}\sum_{i=1}^{n} \tfrac{1}{\sqrt{2\lambda}} \exp(\tfrac{-\sqrt{2}|x-X_i|}{\sqrt{\lambda}}) \le \hat{f}(x) \le \frac{1}{n}\sum_{i=1}^{n} \tfrac{1}{\sqrt{\lambda}} \exp(\tfrac{-2|x-X_i|}{\sqrt{\lambda}}).$$

Positivity of the second component in the second equation is treated in detail by Eggermont and LaRiccia (1999).

The upper bound is a proper density, but the lower bound is a sub-density. One may wonder how a density function can be a upper bound for another density function? The answer lies in the fact that the solution of the maximum penalized likelihood method (5.1) is itself is a sub-density, that is its total mass is always less that unity. Indeed, Eggermont

and LaRiccia (1999) show that for solutions of (5.1),

$$\int (\hat{v}(x))^2 dx = 1 - \lambda \int (\hat{v}_x(x))^2 dx,$$

and thus can be easily renormalized to have mass one.

For the Good and Gaskins' estimator, then, we do not have a exact representation of the penalty estimator in terms of a kernel estimator, and we are led to believe that nonlinearities are likely to render this unlikely in the case of most other penalty methods. Nevertheless, the interplay between the penalty approach and the kernel approach constitute, in our view, a fruitful means of better understanding both methods.

6. Conclusion

We have elaborated some connections between kernel and penalty methods of density estimation, illustrating that exact equivalence can be achieved by adopting a Pearsonian measure of fidelity, or goodness-of-fit, combined with certain $\mathcal{L}_2$ roughness penalties. The quadratic structure of such variational problems leads to exact solutions representable by integral equations and interpretable as kernel estimators. Higher order derivative penalties yield higher-order kernels with their attendant advantages and disadvantages, notably their tendency to deliver negative estimates of the density in the tails. Modification of these penalty problems to impose non-negativity or more exotic properties like log-concavity are quite straightforward. Indeed, we would argue that the virtue of penalty methods generally is their flexibility, the opportunity afforded to tailor both fidelity and penalty contributions to the demands of particular applications. A large class of such problems retains a convenient convex structure that facilitates efficient computations via modern interior point methods. Some further details emphasizing penalized likelihood methods with total variation roughness penalties are available in Koenker and Mizera (2006).

References

Aidu, F. A., and V. N. Vapnik (1989): "Probability Density Estimation via Stochastic Regularization," *Automatic Remote Control*, 50, 84–97.

Boneva, L. I., D. G. Kendall, and I. Stefanov (1971): "Spline Transformation: Three New Diagnostic Aids for the Statistical Data-analyst (with Discussion)," *Journal of the Royal Statistical Society, Series B: Methodological*, 33, 1–70.

Carrier, G. F., and C. E. Pearson (1988): *Partial differential equations*. Academic Press Inc., Boston, MA, second edn., Theory and technique.

Chaudhuri, P., and J. S. Marron (2000): "Scale space view of curve estimation," *The Annals of Statistics*, 28(2), 408–428.

Cox, D. D., and F. O'Sullivan (1990): "Asymptotic analysis of penalized likelihood and related estimators," *Ann. Statist.*, 18(4), 1676–1695.

de Montricher, G. F., R. A. Tapia, and J. R. Thompson (1975): "Nonparametric maximum likelihood estimation of probability densities by penalty function methods," *Ann. Statist.*, 3(6), 1329–1348.

Eggermont, P. P. B., and V. N. LaRiccia (1999): "Optimal convergence rates for Good's nonparametric maximum likelihood density estimator," *Ann. Statist.*, 27(5), 1600–1615.

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

Koenker, R., and I. Mizera (2006): "Density estimation by total variation regularization," in *Advances in statistical Modeling: Essays in Honor of Kjell Doksum*, ed. by V. Nair. World Scientific, Singapore.

Li, Y., and D. Ruppert (2008): "On the Asymptotics of Penalized Splines," *Biometrika*, 95, 415–436.

McOwen, R. (2003): *Partial Differential Equations: Methods and Applications*. Prentice-Hall.

Silverman, B. W. (1982): "On the estimation of a probability density function by the maximum penalized likelihood method," *Ann. Statist.*, 10(3), 795–810.

——— (1984a): "Some remarks on roughness penalty density estimators," in *Limit Theorems in Probability and Statistics*, ed. by P. Révész, vol. 2, pp. 957–980. Colloquia Mathematica Societatis János Bolyai.

——— (1984b): "Spline Smoothing: The Equivalent Variable Kernel Method," *The Annals of Statistics*, 12, 898–916.

Strauss, W. A. (1992): *Partial differential equations*. John Wiley & Sons Inc., New York, An introduction.

Terrell, G. (1990): "Linear Density Estimates," in *Proceedings of the Statistical Computing Section*, pp. 297–302. American Statical Association.

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