Submitted to the Annals of Statistics

OPTIMAL PROBABILITY DENSITY ESTIMATION VIA DIFFUSION EQUATIONS

By John Nixon

Formerly: Agriculture and Agri-Food Canada
E-mail: John.h.Nixon1@gmail.com

An outline of a general strategy is proposed to obtain an estimate of the probability density from a set of data. This method assumes that the estimate is the solution of a diffusion equation with coefficients that can vary continuously on the space of the data and its integrated weighted mean square error is minimised. This formulation is for the one-dimensional case only but seems likely to be generalisable. The mathematical formulation of this optimisation problem requires the use of functional differential calculus and Lagrange multipliers for optimisation of functions with constraints. Surprisingly the resulting equations, together with the use of the adjoint of the diffusion operator, could be reduced to a set of integral equations using the Greens function for the diffusion operator. Further, because of a free parameter involved in the equations that does not affect the result, the Greens function could be replaced by the asymptotic small t WKBJ approximation for it, theoretically with error in the solution of the equations approaching zero as this parameter approaches zero. A problem has so far prevented any valid numerical results being obtained and an error in the calculations is suspected. Please send any comments to me at the email address above. The computer program can be found here.

1. Introduction. The estimation of a one-dimensional probability density $f(x)$ from a set of data $\{X_1, X_2, \ldots X_N\}$ is an old problem with numerous applications in fields as diverse as astronomy and bioinformatics and economics, including estimating the frequencies of rare events.

This fundamental problem may be thought of a simple problem in functional estimation that simultaneously gives estimates of all the function values at an infinite number of points. There are many well known methods available and a great variety of interesting ideas and methods are now under investigation for its solution. These overlapping classes of methods range from the most simple such as the histogram, via kernel density methods [6, 10, 11], to methods involving (1) minimising functionals (e.g. the log-

MSC 2010 subject classifications: Primary 62G07 Secondary 49-xx
Keywords and phrases: Density estimation, Calculus of variations, Fokker Planck equation, Ito Diffusion

likelihood [7], the least squares cross-validation (LCSV) which is a practical version of the integrated mean square error (IMSE)[6], and special cases of the Csiszar Measure which are measures of divergence between two distributions such as the Kullback-Leibler (KL) information [8] or the Pearson χ^2 discrepancy measure [9]), and (2) diffusion equations [3].

Common to all these approaches is the concept of smoothing which means that the density estimation at a point depends on the density in a neighbourhood of that point, and the bandwidth is a measure of the width of this neighbourhood. There is in general a trade-off between bias and variance of an estimate of the density. Over smoothing (i.e. under-fitting) using a large bandwidth reduces variance of the estimated density because an average based on a larger sample of near data points is used, but at the expense of larger bias because of the effect of sampling from a larger region where the underlying density may be different from the point at which it is being estimated. The reverse is obviously true for under-smoothing (i.e. over-fitting) with a small bandwidth. The balance of this trade-off is obviously dependent on the rate at which the density seems to be changing with less variation of density favouring larger bandwidths. In this paper the IMSE will be the criterion to decide what level of fitting is optimal.

Much research has been devoted to how to optimise the choice of bandwidth [6, 10] (because it can be dependent on x) in the context of the popular kernel density methods in which it appears explicitly as a parameter. In these methods each data point corresponds to an additive component of the estimated density (the kernel function) and the criterion for goodness of fit is usually based on the IMSE. Optimising the choice of the kernel function (often taken to be the Gaussian distribution) has also received considerable attention [10].

A fascinating study [3] has been undertaken to find diffusion i.e. Fokker Planck (FP) equations with variable coefficients that generate good estimators of probability densities. This approach deals well with bias near the boundaries for densities that are known to have support in (i.e. have zero value outside) an interval. However, in this method these variable coefficients are fixed arbitrarily rather than being allowed to vary to minimise the IMSE.

The constant coefficient diffusion equation without boundary conditions has Gaussian solutions and is linear allowing additive superposition of solutions, thus simple Gaussian kernel density estimates are solutions of the constant coefficient diffusion equation [3]. Even in the more general variable coefficient case, the solutions integrate to 1 if the initial condition does so, provided the Neumann boundary condition $\frac{\partial f}{\partial x} = 0$ holds at any finite end

points. Thus density estimates based on the FP equation and the Neumann boundary condition generalise the simple kernel density approach. This generalisation allows for the possibility of minimising the IMSE by varying simultaneously both (1) the variable bandwidth (i.e. local adaptivity) and (2) variable shifting (via the drift coefficient obtained from the FP equation) involved in kernel smoothing. These functions are related to the two variable coefficients in the FP equation [3]. Data shifting is also known as data sharpening [5] and was introduced in the context of kernel density methods to offset an effect of data smoothing which is to broaden peaks in the estimated density. This optimisation can be accomplished by choosing both variable coefficents in the FP equation so as to simultaneously minimise the IMSE. Carrying out this optimisation will simultaneously determine what the kernel function should be (including allowing it to vary with the local density if needed) from the solution of the FP equation. In this way, using a diffusion process (actually an Ito diffusion) will generate from every delta function or point mass function in the initial empirical density, a distribution that spreads out with time, which approximates a Gaussian distribution and is exactly so if the diffusion equation has constant coefficients.

In this paper this problem is formulated as an optimisation problem for the weighted IMSE subject to constraints imposed by the diffusion model without boundary conditions. This optimisation problem was reduced to a set of integrodifferential equations using the calculus of variations with the notation that I developed for functional derivatives [1, 2] in connection with applications in statistical physics, and so the paper may be hard to read without this mathematical background. This notation reduces the taking of functional derivatives of the most complex expressions to a mechanical procedure in analogy with ordinary differentiation. This work therefore builds on the work of [3] and avoids their complex method of bandwidth determination. The weighting in the IMSE was introduced (treated as an externally given function that can be chosen arbitrarily) to allow for concentrating more heavily on a particular small region of x space when minimising the IMSE. In section 2 the MSE is introduced, its squared bias and variance components and also the IMSE. The IMSE is shown to have a much simpler expression if the estimator for the density is assumed to be a sum of terms originating from each data point. In this case it is shown that the variational problem for the IMSE only has a trivial solution. It is also shown that the IMSE has a minimum but no maximum and the general diffusion model is introduced that restricts the above terms to be solutions of an FP equation. In section 3 the Langrange equations are written down for the minimisation of the IMSE subject to the diffusion model as a constraint,

and formal solutions of them are found using the properties of the adjoint of the diffusion operator, and simplification using the Chapman Kolmogoroff equation. This involves a free parameter in the equations, K, that does not affect the solution. In section 4, K is used to allow the substitution of the exact solution of the diffusion equation by its asymptotic small t (WKBJ) approximation without any error asymptotically as $K \to 0$. In section 5 the derived numerical scheme is formulated, and some concluding comments are given including a description of the problem of numerical integration of a ratio where the denominator passes through 0, which is needed to complete the numerical algorithm for density estimation.

2. The IMSE, its minimisation and the diffusion model. For estimating the unknown density $f(x)$ for $x \in \mathbb{R}$ from an independently and identically distributed (IID) sample of N observed values $X_1 \dots X_N$, suppose an estimator for $f(x)$ is $\hat{f}(x) = \hat{f}(x, X_1, \dots X_N)$ for the density $f(x)$. For fixed $\hat{f}$, the average over samples is

$$E[\hat{f}(x)] = \int dX_1 f(X_1) \dots \int dX_N f(X_N) G(x, X_1, \dots X_N). \tag{1}$$

This equation also defines the meaning of $E[\dots]$ and allows a compact representation without the use of multiple integration signs. The integrals over the space of the data will be indicated like this without limits. The mean squared error (MSE) is defined as

$$E[(\hat{f}(x) - f(x))^2] = E[(\hat{f}(x) - E[\hat{f}(x)])^2] + (E\hat{f}(x) - f(x))^2. \tag{2}$$

which follows after writing $\hat{f}(x) - f(x) = (\hat{f}(x) - E\hat{f}(x)) + (E\hat{f}(x) - f(x))$, squaring and doing the expectation term by term (from which the cross term has expectation equal to 0). This uses the linearity of E and that $E\hat{f}(x)$ and $f(x)$ are not dependent on a sample so can be factored out of an expectation. The first term of Equation (2) further simplifies so Equation (2) becomes

$$E[(\hat{f}(x) - f(x))^2] = E[\hat{f}^2(x)] - (E[\hat{f}(x)])^2 + (E\hat{f}(x) - f(x))^2. \tag{3}$$

Thus the MSE is composed of terms representing the variance and squared bias of the estimate and is dependent on the point at which the estimate is being made, the unknown true density and the sample size, but independent of the sample $\{X_1, X_2, \dots X_N\}$ from which the density is estimated because it is an average of all possible samples. Using the notation

$$I_i(x) = E\left[\hat{f}(x)^i\right] \text{ for } i = 1, 2 \tag{4}$$

the MSE simplifies to $I_2(x)-2I_1(x)f(x)+f^2(x)$. The integral of this quantity over the whole range of x, the integrated mean square error (IMSE), would be an ideal candidate for minimisation in a variational argument to determine the best estimate of the true density from a set data. The integrated mean square error (IMSE) is defined as

$$M = \int dx w(x) \left[I_2(x) - 2I_1(x)f(x) + f^2(x) \right] \tag{5}$$

which is the total measure of error for the estimator $\hat{f}(x)$ for $f(x)$, and the weight function $w(x) > 0$ for all x. Clearly $M > 0$ so M must have a minimum value. If the assumption is made that the separate data points contribute additively to the estimate of $f(x)$, we can write

$$\hat{f}(x, X_1, \ldots X_N) = \frac{1}{N} \sum_{i=1}^{N} G(x, X_i, t_0). \tag{6}$$

Using (6) the general expression for the MSE can be simplified. First note that

$$\begin{aligned} I_1(x) &= \frac{1}{N} \sum_{i=1}^{N} \int dX_1 f(X_1) \ldots \int dX_N f(X_N) G(x, X_i, 1) \\ &= \frac{1}{N} \sum_{i=1}^{N} \int dX_i f(X_i) G(x, X_i, 1) = \int dX f(X) G(x, X, 1). \end{aligned} \tag{7}$$

Similarly

$$I_2(x) = \int dX_1 f(X_1) \ldots \int dX_N f(X_N) \frac{1}{N} \sum_{i=1}^{N} G(x, X_i, 1) \frac{1}{N} \sum_{j=1}^{N} G(x, X_j, 1)$$

can be simplified term by term, factorising out the integrals that are 1 and splitting the double sum into $i \neq j$ and $i = j$ sums. The expression simplifies to

$$I_2(x) = \frac{N(N-1)}{N^2} \left(\int dX f(X) G(x, X, 1) \right)^2 + \frac{1}{N} \int dX f(X) G^2(x, X, 1). \tag{8}$$

Therefore the MSE becomes

$$\begin{aligned} &\frac{1}{N} \int dX f(X) G^2(x, X, 1) + \left(1 - \frac{1}{N} \right) \left(\int dX f(X) G(x, X, 1) \right)^2 \\ &\qquad -2f(x) \int dX f(X) G(x, X, 1) + f^2(x). \end{aligned} \tag{9}$$

Hence the IMSE can be written as

$$\begin{aligned} M = & \frac{1}{N}\int dXf(X)\int dyw(y)G^2(y,X,1) \\ & + \left(1-\frac{1}{N}\right)\int dyw(y)\left(\int dXf(X)G(y,X,1)\right)^2 \\ & - 2\int dXf(X)\int dyw(y)f(y)G(y,X,1) + \int dyw(y)f^2(y). \end{aligned} \tag{10}$$

Calculating the functional derivative formally [1] gives

$$\begin{aligned} \frac{\delta M}{\delta G(x,Y,s)} = & \frac{1}{N}\int dXf(X)\int dyw(y)2G(y,X,1)\delta(y-x)\delta(X-Y)\delta(1-s)+ \\ & \left(1-\frac{1}{N}\right)\int dyw(y)\left[2\int dXf(X)G(y,X,1)\times\right. \\ & \left.\int dXf(X)\delta(y-x)\delta(X-Y)\delta(1-s)\right] \\ & - 2\int dXf(X)\int dyf(y)w(y)\delta(y-x)\delta(X-Y)\delta(1-s) \end{aligned} \tag{11}$$

which simplifies to

$$\begin{aligned} \frac{\delta M}{\delta G(x,Y,s)} = & \delta(1-s)f(Y)w(x)\left[\frac{2}{N}G(x,Y,1)+\right. \\ & \left. 2\left(1-\frac{1}{N}\right)\int dXf(X)G(x,X,1) - 2f(x)\right]. \end{aligned} \tag{12}$$

Hence, just using the linearity (6), equating (12) directly to zero gives

$$H(x,Y) = Nf(x) + (1-N)\int dXf(X)H(x,X) \tag{13}$$

where $H(x,Y) = G(x,Y,1)$. Now (13) shows that H is independent of its second argument which implies that $G(x,Y,1) = H(x) = f(x)$. Then from (6) it follows that trivially

$$\hat{f}(x,X_1,\dots X_N) = f(x). \tag{14}$$

This shows that any point $\hat{f}(x)$ that is a stationary point of M (this includes all local and global maximum or minimum points) must have $\hat{f}(x) = f(x)$ so there is only one such point. It is clear that M can always be increased so it has no maximum, for example let

$$G(x,Y,1) = \begin{cases} 1/\delta & |x-x^*| < \delta/2 \\ 0 & \text{Otherwise} \end{cases} \tag{15}$$

which is independent of its second argument (note this would give a silly choice of $\hat{f}$ because it does not depend on the sample). Then from (10)

$$M \approx \frac{1}{\delta} \times \left[\frac{w(x^*)}{N} + \left(1 - \frac{1}{N}\right) w(x^*) \right] - 2w(x^*)f(x^*) + \int dy w(y) f^2(y)$$

$$= \frac{1}{\delta} - 2w(x^*)f(x^*) + \int dy w(y) f^2(y) \to \infty \tag{16}$$

as $\delta \to 0$. Thus because $M > 0$, M has a minimum but no maximum. Therefore the unique stationary point cannot be a local maximum or local saddle point (then there would be another lower local minimum point that is stationary), so it is the unique global minimum point. This shows that the best estimator of $f(x)$ subject only to linearity (6) is $f(x)$ itself, and the result is independent of the sample. This establishes the validity of the IMSE criterion but seems analogous to fitting a saturated model to data giving an exact fit. The error is zero but the model has no predictive value so it is useless in applications. Therefore following Botev et al. [3] a model density is needed allowing the IMSE to be calculated, and containing parameters that can be varied to minimise the IMSE. The density estimate is the solution of the FP equation at $t = t_0$. Rather than reverting to explicit parameterised models, the model I propose here is the FP equation

$$\frac{\partial G}{\partial t} = \frac{1}{2} \frac{\partial}{\partial x} \left(a(x) \frac{\partial}{\partial x} \big(b(x) G(x,t) \big) \right) \tag{17}$$

containing the two variable coefficient functions of the independent variable as parameters to be fitted that are related to the drift velocity and variance of the corresponding random process. The initial condition is $G(x,0) = \frac{1}{N} \sum_{i=1}^{N} \delta(x - X_i)$. Here $a(.)$ and $b(.)$ have the same sign for a solution with $t > 0$ [Botev sec. 3]. The solution is unique [maximum principle], and because of the linearity of this equation, the solutions are additive, where $G(x, X, t_0)$ is the solution of (17) satisfying $G(x, X, 0) = \delta(x - X)$. Without loss of generality t_0 can be chosen to be 1 because any scaling of t can be accommodated by multiplying the as yet arbitrary function $a(x)$ by a constant. The defining equations for the Green's function $G(x, X, t)$ can now be stated as the following initial value problem

$$\frac{\partial}{\partial t} G(x, X, t) = \frac{1}{2} \frac{\partial}{\partial x} \left(a(x) \frac{\partial}{\partial x} \left(b(x) G(x, X, t) \right) \right) \tag{18}$$

$$G(x, X, 0) = \delta(x - X) \tag{19}$$

to which the condition

$$a(x_0)b(x_0) = K \tag{20}$$

is added to control an overall multiplicative constant on the right hand side of (18) because this controls the rate of evolution of G with t from 0 to 1. The value x_0 is arbitrary but can be related to the data in some way e.g. its mean.

I now propose to minimise the IMSE subject to the three conditions (18), (19) and (20) on the form of G, again using the calculus of variations in the hope of determining functions $a(x)$ and $b(x)$ and the value K thereby determining the additive model (6) that minimises M, where G satisfies the FP equation (18) with initial condition (19). For numerical purposes clearly Equation (6) will be needed as the new estimate of f from the data sample has not up to this point been used. This suggests an iterative procedure will be needed when implementing this scheme.

3. Formulating and simplifying the Langrange equations. Minimisation of M subject to the constraints can be done by using Langrange's method of undetermined multipliers as follows. First form the quantity

$$\begin{aligned} W = M + \int dy \int dY \int_0^1 dt\, \lambda_1(y,Y,t) \Big\{ \frac{\partial G(y,Y,t)}{\partial t} - \\ \frac{1}{2}\frac{\partial}{\partial y}\left[a(y) \frac{\partial}{\partial y} \left(b(y) G(y,Y,t) \right) \right] \Big\} + \\ \int dy \int dY \lambda_2(y,Y) \left[G(y,Y,0) - \delta(y-Y) \right] + \lambda_3 (a(x_0)b(x_0) - K). \end{aligned} \tag{21}$$

The variables $G(.,.,.)$, $a(.)$ and $b(.)$ and K are to be varied simultaneously to as to satisfy the stationarity condition i.e. $\frac{\delta W}{\delta G(x,Y,s)}$, $\frac{\delta W}{\delta a(x)}$, $\frac{\delta W}{\delta b(x)}$ and $\frac{\partial W}{\partial K}$ are all zero. The λ's are fixed unknown quantities (i.e. unknown functions dependent only on the stated variables). The functions $G()$, $a(.)$ and $b(.)$ are treated as independent (function) variables varying over spaces of functions, and K is an independent numerical variable, so as each variable is varied, all the others are treated as constant while taking all the derivatives. The resulting equations are to be solved simultaneously for $G()$,$a(.)$, $b(.)$ and K as well as the λ functions. The formulation of this is analogous to the finite dimensional case of the use of Langrange multipliers to optimise a function subject to constraints. The constraint equations are now (18),(19) and (20) and form a 3-parameter family, a 2-parameter family, and a single equation respectively. Firstly, the derivative w.r.t. K implies

$$\lambda_3 = 0. \tag{22}$$

Calculating the functional derivatives using all the usual properties of derivatives is straightforward though a little tedious. The integrations can be treated like sums over dummy variables and the derivatives are treated like differences so the functional derivatives can go inside all these operations and the functional deriviative of a function w.r.t. itself at another argument is a delta function. This together with (22) leads to the three remaining equations:

$$\begin{aligned} 0 =& \frac{\delta W}{\delta G(x,X,s)} = \frac{\delta M}{\delta G(x,X,s)} + \\ & \int dy \int dY \int_0^1 dt \lambda_1(y,Y,t) \times \left\{ \frac{\partial}{\partial t} \delta(y-x)\delta(Y-X)\delta(t-s) - \right. \\ & \left. \frac{1}{2}\frac{\partial}{\partial y}\left[a(y) \frac{\partial}{\partial y} \left(b(y)\delta(y-x)\delta(Y-X)\delta(t-s) \right) \right] \right\} + \\ & \int dy \int dY \lambda_2(y,Y)\delta(y-x)\delta(Y-X)\delta(-s) = 0, \end{aligned} \tag{23}$$

$$0 = \frac{\delta W}{\delta a(x)} = -\frac{1}{2} \int dy \int dY \int_0^1 dt \, \lambda_1(y,Y,t) \frac{\partial}{\partial y} \left[\delta(y-x) \frac{\partial}{\partial y} \left(b(y) G(y,Y,t) \right) \right] \tag{24}$$

and

$$0 = \frac{\delta W}{\delta b(x)} = -\frac{1}{2} \int dy \int dY \int_0^1 dt \, \lambda_1(y,Y,t) \frac{\partial}{\partial y} \left[a(y) \frac{\partial}{\partial y} \left(\delta(y-x) G(y,Y,t) \right) \right]. \tag{25}$$

In Equations (72),(73) and (74) the δ functions and their derivatives have to be integrated over. The simplest way to do this is to do the integrations by parts where one of the factors is the derivative of a term, and the terms evaluated at the limits at $\pm\infty$ are zero. Sometimes this has to be done twice. Finally all the delta functions can be integrated out. Then Equation 72 becomes

$$\begin{aligned} & \frac{\delta M}{\delta G(x,X,s)} + \delta(1-s)\lambda_1(x,X,1) - \delta(s)\lambda_1(x,X,0) - \frac{\partial}{\partial s}\lambda_1(x,X,s) \\ & \qquad -\frac{1}{2}\frac{\partial}{\partial x}\left[a(x) \frac{\partial}{\partial x} \lambda_1(x,X,s) \right] . b(x) + \lambda_2(x,X)\delta(s) = 0 \end{aligned} \tag{26}$$

Similarly Equation 73 becomes

$$\frac{1}{2} \int dY \int_0^1 dt \frac{\partial}{\partial x} \left(b(x) G(x,Y,t) \right) \frac{\partial \lambda_1(x,Y,t)}{\partial x} = 0 \tag{27}$$

and finally in the same way it follows from Equation 74 that

$$-\frac{1}{2}\int dY\int_0^1 dt\, G(x,Y,t)\frac{\partial}{\partial x}\left(a(x)\frac{\partial\lambda_1(x,Y,t)}{\partial x}\right)=0 \tag{28}$$

Equation (26) needs to be combined with Equation (12). This equation then has δ function singularities at $s=0$ and at $s=1$ and other terms varying smoothly with s. Equating this to zero requires that each of these terms is separately is equal to zero. Then it follows that

$$\begin{aligned}&\lambda_1(x,X,1)+f(X)w(x)\left[\frac{2}{N}G(x,X,1)\right.\\ &\left.+2\left(1-\frac{1}{N}\right)\int dY f(Y)G(x,Y,1)-2f(x)\right]=0 \qquad (29)\\ &-\lambda_1(x,X,0)+\lambda_2(x,X)=0 \qquad (30)\\ &\text{and}\quad -\frac{\partial\lambda_1}{\partial s}(x,X,s)-\frac{1}{2}\frac{\partial}{\partial x}\left(a(x)\frac{\partial}{\partial x}\lambda_1(x,X,s)\right)b(x)=0. \qquad (31)\end{aligned}$$

From these equations at first sight it appears that λ_2, as the initial condition for λ_1, could be any function because Equation (31) is the evolution equation for λ_1 with an initial condition at $s=0$ that determines λ_1 at $s=1$ where it is related to $G(x,Y,1)$ via (29). This is wrong because of a sign change in Equation (31) when compared with Equation (18) which shows that actually λ_1 evolves backwards from its value at $s=1$ to its value at $s=0$ because the solution of Equation (18) is only valid for $t\geq 0$. The result of the analysis with Langrage multipliers is the equations (73) to (31) together with the original equations (18), (19) and (20). From its role in the equations as a scale factor in the r.h.s. of Equation (18) K appears to be basically a bandwidth parameter because it effectively alters the amount of time evolution from the initial empirical density. It is interesting that the optimisation method described here does not determine K because the only term containing it is multiplied by λ_3 which is zero. Therefore its value cannot affect the final result. In Equation (31) put $s=1-t$ and $\lambda_1(x,X,s)=\phi(X,x,t)$ then Equation (31) becomes

$$\frac{\partial}{\partial t}\phi(X,x,t)-\frac{1}{2}\frac{\partial}{\partial x}\left(a(x)\frac{\partial}{\partial x}\phi(X,x,t)\right)b(x)=0, \tag{32}$$

which is to be solved with the intial condition

$$\phi(X,x,0)=\lambda_1(x,X,1). \tag{33}$$

Equation (32) is obviously related to Equation (18). To explain the connection, consider the following differential operator

$$S[f(x)] = a(x)\frac{\partial}{\partial x}\left(b(x)\frac{\partial}{\partial x}\left(c(x)f(x)\right)\right) \tag{34}$$

Then the following chain of equalities holds:

$$\begin{aligned}
S[\delta(x-y)] &= a(x)\frac{\partial}{\partial x}\left(b(x)\frac{\partial}{\partial x}\left(c(x)\delta(x-y)\right)\right)\\
&= a(x)\frac{\partial}{\partial x}\left(b(x)\frac{\partial}{\partial x}\left(c(y)\delta(x-y)\right)\right)\\
&= c(y)a(x)\frac{\partial}{\partial x}\left(b(x)\frac{\partial}{\partial x}\delta(x-y)\right)\\
&= -c(y)a(x)\frac{\partial}{\partial x}\left(b(x)\frac{\partial}{\partial y}\delta(x-y)\right)\\
&= -c(y)\frac{\partial}{\partial y}\left(a(x)\frac{\partial}{\partial x}\left(b(x)\delta(x-y)\right)\right)\\
&= -c(y)\frac{\partial}{\partial y}\left(a(x)\frac{\partial}{\partial x}\left(b(y)\delta(x-y)\right)\right)\\
&= -c(y)\frac{\partial}{\partial y}\left(b(y)a(x)\frac{\partial}{\partial x}\delta(x-y)\right)\\
&= c(y)\frac{\partial}{\partial y}\left(b(y)a(x)\frac{\partial}{\partial y}\delta(x-y)\right)\\
&= c(y)\frac{\partial}{\partial y}\left(b(y)\frac{\partial}{\partial y}\left(a(x)\delta(x-y)\right)\right)\\
&= c(y)\frac{\partial}{\partial y}\left(b(y)\frac{\partial}{\partial y}\left(a(y)\delta(x-y)\right)\right) \qquad (35)
\end{aligned}$$

Therefore

$$S[\delta(x-y)] = S^*[\delta(x-y)] \tag{36}$$

where the adjoint of the operator S denoted by S^* is defined by

$$S^*[f(y)] = c(y)\frac{\partial}{\partial y}\left(b(y)\frac{\partial}{\partial y}\left(a(y)f(y)\right)\right). \tag{37}$$

This general result makes it clear that $[S^*]^* = S$. From (37) it follows that if the operator L is given by

$$L[f(x)] = \frac{1}{2}\frac{\partial}{\partial x}\left[a(x)\frac{\partial}{\partial x}\left(b(x)f(x)\right)\right], \tag{38}$$

then the adjoint of L is given by

$$L^*[g(y)] = \frac{1}{2}b(y)\frac{\partial}{\partial y}\left[a(y)\frac{\partial g(y)}{\partial y}\right]. \tag{39}$$

Because L and L^* involve different variables for differentiation, the operations of L and L^* can be reversed i.e. $LL^* - L^*L = [L, L^*] \equiv 0$. For the solution $G(x, y, t)$ of $\frac{\partial G}{\partial t} = LG$ we have $L = \frac{\partial}{\partial t}$. Therefore for small δt

$$\begin{aligned}(L - L^*)G(x, y, t + \delta t) &= (L - L^*)[\delta t LG(x, y, t) + G(x, y, t)] \\ &= \delta t L(L - L^*)G(x, y, t) + (L - L^*)G(x, y, t) \\ &= 0 \end{aligned} \tag{40}$$

and for all $x, y \in \mathbb{R}$ provided $(L - L^*)G(x, y, t) = 0$. Therefore since $(L - L^*)G(x, y, t) = 0$ is true from (36) at $t = 0$, it must be true for all $t > 0$ so the equations $\frac{\partial G}{\partial t} = LG(x, y, t)$ and $\frac{\partial G}{\partial t} = L^*G(x, y, t)$ both have the same solution that satisfies the initial condition (19). Another way of describing this is to say that provided G satisfies (19), in the slices for which y is constant there are a set of 2D PDE problems $\frac{\partial G}{\partial t} = LG$ that define the same solution $G(x, y, t)$ in 3 dimensions as the set of 2D PDE problems $\frac{\partial G}{\partial t} = L^*G$ in the slices for which x is constant. Applying this to the solution of Equation (32), notice that this equation is the adjoint of Equation (18), which can be written as

$$\frac{\partial \phi}{\partial t} = \frac{1}{2}\frac{\partial}{\partial X}\left(a(X)\frac{\partial}{\partial X}\left(b(X)\phi(X, x, t)\right)\right). \tag{41}$$

Here x is a free parameter. Thus the solution of (32) with the initial condition

$$\phi(X, x, 0) = \delta(X - x), \tag{42}$$

is the same as the solution of (41) with initial condition (42) and will be denoted by $G(X, x, t)$. The solution of (41) with the general initial condition

$$\phi(X, x, 0) = g(X) \tag{43}$$

is by linearity,

$$\phi(X, t) = \int du G(X, u, t)g(u) \tag{44}$$

because

$$\phi(X, x, 0) = g(X) = \int du \delta(X - u)g(u). \tag{45}$$

Therefore by comparing (43) with (62), the solution of (41) with initial condition (62) is seen to be

$$\lambda_1(x, X, s) = \phi(X, x, 1-s) = \int duG(X, u, 1-s)\lambda_1(x, u, 1). \tag{46}$$

Substituting for λ_1 using the R.H.S. of (46) into (27) gives a complex expression which can be simplified by (1) rearranging it so as to bring the integration over u outside, the integration over Y inside, (2) using the Chapman-Kolmogoroff equation

$$\int dyG(x, y, t_1)G(y, z, t_2) = G(x, z, t_1 + t_2) \tag{47}$$

and (3) doing the trivial integration over t, and assuming integration and differentiation can be exchanged. This gives

$$\int du \frac{\partial \lambda_1(x, u, 1)}{\partial x} \frac{\partial}{\partial x} \left(b(x)G(x, u, 1)\right) = 0. \tag{48}$$

Substituting in λ_1 using (29) gives

$$\int duf(u) \frac{\partial}{\partial x} \left(b(x)G(x, u, 1)\right) \alpha(x, u) = 0. \tag{49}$$

Likewise substituting for λ_1 using (46) into (28) gives

$$\int duG(x, u, 1) \frac{\partial}{\partial x} \left(a(x) \frac{\partial}{\partial x} \lambda_1(x, u, 1)\right) = 0. \tag{50}$$

Now substituting for λ_1 using (29) gives

$$\int duf(u)G(x, u, 1) \frac{\partial}{\partial x} \left(a(x)\alpha(x, u)\right) = 0. \tag{51}$$

In equations (49) and (51)

$$\alpha(x, u) = \frac{\partial}{\partial x} \left(\beta(x, u)w(x)\right) \tag{52}$$

and $\beta(x, u)$ is defined by

$$\beta(x, u) = \frac{1}{N} G(x, u, 1) + \left(1 - \frac{1}{N}\right) \int dzf(z)G(x, z, 1) - f(x). \tag{53}$$

This expression for $\beta(x,u)$ can be simplified by the substitution for f i.e.

$$f(x) = \frac{1}{N}\sum_{i=1}^{N} G(x, X_i, 1) \tag{54}$$

and the Chapman-Kolmogorov equation to obtain

$$\beta(x,u) = \frac{1}{N}\left(G(x,u,1) + \left(1 - \frac{1}{N}\right)\sum_{i=1}^{N} G(x, X_i, 2)\right) - f(x). \tag{55}$$

Now equations (49) and (51) are first order with respect to $b(.)$ and $a(.)$ that are both positive everywhere, therefore they have solutions as the following integrals respectively:

$$b(x) = b(x_0)\exp\left\{-P.V.\int_{x_0}^{x} ds \frac{\int du \frac{\partial G}{\partial s}(s,u,1) f(u)\alpha(s,u)}{D(s)}\right\} \tag{56}$$

and

$$a(x) = a(x_0)\exp\left\{-P.V.\int_{x_0}^{x} ds \frac{\int du G(s,u,1) f(u)\frac{\partial \alpha}{\partial s}(s,u)}{D(s)}\right\} \tag{57}$$

where

$$D(s) = \int du G(s,u,1) f(u)\alpha(s,u) \tag{58}$$

and the initials P.V. indicate the Cauchy principal value which is required because preliminary results show that $D(x)$ can pass through 0 and so the integrals would in this case be otherwise undefined. The principal value removes from the range of integration a small region $\{x : |x - x_s| < \epsilon\}$ symmetrically placed about each singular point x_s and then passes to the limit $\epsilon \to 0$. Notice that the numerators of the internal integrals in (56) and (57) add to $\frac{dD(s)}{ds}$. A consequence of this is that using (20) likewise

$$a(x)b(x) = K\exp\left\{-P.V.\int_{x_0}^{x} ds \frac{\frac{d}{ds}D(s)}{D(s)}\right\} = K\left|\frac{D(x_0)}{D(x)}\right|. \tag{59}$$

Therefore it is unnecessary to compute $\frac{\partial \alpha}{\partial s}$ and once $b(x)$ has been computed, it is faster to compute $a(x)$ using (59) rather than using (57).

4. The use of the WKBJ approximation for small K. To complete the iterative scheme, a method is needed for calculating the Greens function G numerically. This can be done using the fact that K is completely arbitrary and can therefore be made as close to zero as needed without theoretically introducing any error in the solution of the complete set of equations determining $f(.)$,$a(.)$,$b(.)$ etc.. If the new variable $t^+ = Kt$ is introduced, and G^+ is introduced by

$$G^+(x, X, t^+) = G(x, X, t) \tag{60}$$

then it is easy to show that G^+ satisfies the equations

$$\frac{\partial}{\partial t^+} G^+(x, X, t^+) = \frac{1}{2}\frac{\partial}{\partial x}\left(a(x)\frac{\partial}{\partial x}\left(b^+(x)G^+(x, X, t^+)\right)\right) \tag{61}$$

$$G^+(x, X, 0) = \delta(x - X) \tag{62}$$

where $b^+(x) = b(x)/K$ and $a(x_0)b^+(x_0) = 1$. Now the small t^+ expansion (Botev et al.[3] Lemma 1) of the solution of (61) with (62) gives

$$G^+(x, u, t^+) = \frac{|b^+(u)|^{1/4}}{\sqrt{2\pi t^+}\,|b^{+3}(x)a(x)a(u)|^{1/4}} \exp\left\{-\frac{1}{2t^+}\left[\int_u^x ds\,|a(s)b^+(s)|^{-1/2}\right]^2\right\}, \tag{63}$$

which is asymptotically exact as $t^+ \to 0$ which implies $K \to 0$ when t is fixed. Expressing this in terms of G,a,b and t it follows that the factors of K cancel and

$$G(x, u, t) = \frac{|b(u)|^{1/4}}{\sqrt{2\pi t}\,|b^3(x)a(x)a(u)|^{1/4}} \exp\left\{-\frac{1}{2t}\left[\int_u^x ds\,|a(s)b(s)|^{-1/2}\right]^2\right\}. \tag{64}$$

Because the equations other than (64) are exact whatever the value of K is, the system of equations for $f(.)$ will be asymptotically exact as $K \to 0$ for any fixed t and a suitable small value of K should be used when implementing this numerically.

5. Summary and conclusions. The above reasoning indicates that the calculation of the estimate of the probability density $f(.)$ from a sample of data $\mathbf{X} = \{X_1, \ldots X_N\}$ that is optimal in the sense that the IMSE with weight function $w(.)$ is minimised subject to $\hat{f}(.)$ satisfying the general second order diffusion equation could be done using an iterative method to simultaneously estimate $a, b, G, \hat{f}, \beta, \alpha, D$. Replacing the variables $a, b,$ and G by their logarithms (denoted by c, d and H respectively) because these values can get extremely small leading to numerical errors simplifies the equations and are written here again in this algorithm for convenience.

1. Choose K.
2. Choose an interval I containing all the data.
3. Choose initial estimates of $c(.)$ and $d(.)$ such that $c(x_0) = 0$, $d(x_0) = \ln(K)$ where $x_0 \in I$ for example x_0 could be the median of $\mathbf{X}$, and $c(.)$ and $d(.)$ could be constant initially.
4. Calculate $H(x, Y, t)$ in $I \times I \times \{1, 2\}$ (or at least that part of it where G is not extremely small) using the WKBJ approximation (64).
(65)
$$H(x,u,t) = \frac{d(u)}{4} - \frac{\ln(2\pi t)}{2} - \frac{3d(x) + c(x) + c(u)}{4} - \frac{1}{2t}\left[\int_u^x ds \exp\left(-\frac{c(s) + d(s)}{2}\right)\right]^2$$
5. Calculate $G(x, u, t) = \exp(H(x, u, t))$.
6. From Equation (54) calculate $\hat{f}(.)$:
$$f(x) = \frac{1}{N}\sum_{i=1}^{N} G(x, X_i, 1). \tag{66}$$
This is where there is input from the data $\mathbf{X}$.
7. From Equation (55) calculate $\beta(.,.)$:
$$\beta(x,u) = \frac{1}{N}\left(G(x,u,1) + \left(1 - \frac{1}{N}\right)\sum_{i=1}^{N} G(x, X_i, 2)\right) - f(x). \tag{67}$$
8. From Equation (52), calculate $\alpha(.,.)$:
$$\alpha(x,u) = \frac{\partial}{\partial x}(\beta(x,u)w(x)) \tag{68}$$
9. From Equation (58) calculate $D(.)$:
$$D(s) = \int du G(s,u,1) f(u) \alpha(s,u) \tag{69}$$
10. From Equation (56) calculate $d(.)$:
$$d(x) = \ln(K) - P.V. \int_{x_0}^{x} ds \frac{\int du G(s,u,1)\frac{\partial H}{\partial s}(s,u,1) f(u)\alpha(s,u)}{D(s)} \tag{70}$$
11. From Equation (59) calculate $c(.)$:
$$c(x) = \ln(K) + \ln|D(x_0)| - \ln|D(x)| - d(x) \tag{71}$$
12. Repeat from step 4 until convergence of $c(.)$ and $d(.)$ or $\hat{f}(.)$.

In these equations of course $f(.)$ is replaced by its latest estimate $\hat{f}$.

Since G is a Green's function, its value can never be 0 or ∞ when $t > 0$. This rules out H from being $\pm\infty$ therefore $c(.)$ and $d(.)$ can likewise not be $\pm\infty$. (For example only simple poles in $c(.)$ or $d(.)$ would in equation (65) imply the last term has essential singularities there so $c(.)$ and/or $d(.)$ would do as well in contradiction to the assumption). This in turn implies $D(s) \neq 0$ for any s in I otherwise setting x to this value in (70) would give an infinite value for $d(x)$ at a logarithmic singularity there. Indeed forcing $D(.)$ to be nonzero e.g. the constant 1 gives a numerical algorithm that rapidly converges to a good looking solution which is however very dependent on the value of K.

At the moment there are two major problems with the density estimation method in this paper (based on a dataset of two points, the simplest case for which this method should work) with constant initial estimates of $c(.)$ and $d(.)$ (implying an approximate Gaussian kernel density estimate of the probability density of the data) and a finite interval I chosen containing the data:

- the calculated value of $D(.)$ appears to always cross zero in disagreement with the above argument, and
- the result is very K dependent especially as $K \to 0$ with divergence of the algorithm seen if K is sufficiently small.

These results were obtained with the formula in the appendix to calculate the (70) as principal value integral that allows the denominator D to cross zero in the hope that the final $D(.)$ converged to would not cross zero in I. A modified algorithm was also tried in which the singularities were removed by simply ignoring the last term in equation(84). In the final result they should be absent to prevent G from being 0 or ∞ at any point. This is rationale for removing them: if the resulting algorithm converges and the denominator $D(s)$ generated has no zero crossing points then the result is also a solution of the original equations. The result again had $D(s)$ crossing zero demonstrating the inconsistency. My belief at the moment is that there is an error somewhere in this paper or in the algorithm here that if corrected would give an efficient algorithm to solve this problem numerically. One thing that have not investigated yet is taking the WKBJ approximation to higher order. However I doubt whether this would qualitatively alter the results, which would be needed to make this work.

To do: demonstrate both translation and scale invariance in this theory of density estimation. A nonlinear change of x variable gives rise to an extra x dependent multiplying factor in Equation (17). Is it possible to combine

density estimates from subsets of the data? For example two isolated clusters of data points would be expected to have a density estimate close to the average of the density estimates associated with the separate clusters, weighted by the number of data points in each.

In this paper I concentrated on the one-dimensional case only but the method should be extendable to many dimensions. Theoretically it would also be interesting to know how much of this could be generalised to higher order PDEs for the density estimate, and if this is possible, what order of PDE would be most appropriate in any given instance. Also extending this to the case where the data points are known to fall into a finite interval eg $[0, 1]$, which would involve using the Neumann boundary condition, would allow this method to be used for estimating probability densities for sets of p-values which could be used in multiple hypothesis testing [4]. This is likely to be most useful when there are a large number of statistical tests to be carried out and the amount of information in each test is limited, so the p-values do not cluster much near zero, otherwise the p-value density is likely to be a monotone decreasing function of p-value and then the sorting the observations by p-value density highest first is equivalent to sorting the observations by p-value lowest first, and trivially the most significant sets of observations should then be decided simply on the basis of p-value [4].

References.

[1] Nixon, John H. (1987). A new approach to the calculation of thermodynamics and structure for classical one-dimensional systems with pairwise additive potentials in an external field. *J. Phys. A* **20** 651–667. MR880816

[2] Nixon, John H. (1994). Minimisation of dimension for functional-differential equations and the thermodynamics of the classical one-dimensional fluid. *J. Phys. A* **27** 1407–1426. MR1279515

[3] Botev, Z. I. and Grotowski, J. F. and Kroese, D. P. (2010). Kernel density estimation via diffusion. *Ann. Statist.* **38** 2916-2957. MR2722460

[4] Nixon, John H. (2013). On multiple hypothesis testing maximizing the average power. *Internat. J. Statist. Probab.* **2** 112–135. International Journal of Statistics and Probability; Vol. 2, No. 2; 2013 pp112-135 doi:10.5539/ijsp.v2n2p112 URL: http://dx.doi.org/10.5539/ijsp.v2n2p112

[5] Hall, Peter and Minnotte, Michael C. (2002). High order data sharpening for density estimation. *J. Roy. Statist. Soc. Ser. B* **64** 141–157. MR1883130

[6] Sheather, Simon J. (2004). Density estimation. *Statist. Sci.* **19** 588–597. MR2185580

[7] Tabak, E.G. and Turner, Cristina V. (2013). A family of nonparametric density estimation algorithms. *Commun. Pure Appl. Math.* **66** 145–164. MR2999294

[8] Chang, Lin and Chong-xiu, Yu (2013). Modified self-organizing mixture network for probability density estimation and classification. *The 2013 International Joint Conference on Neural Networks (IJCNN) 4-9 August 2013, pages 1-6* URL: http://dx.doi.org/10.1109/IJCNN.2013.6706730

[9] Botev, Zdravko I. and Kroese, Dirk P. (2009). The Generalized Cross Entropy

Method, with Applications to Probability Density Estimation. *Methodol. Comput. Appl. Probab.* **13** 1–27. URL: http://dx.doi.org/10.1007/s11009-009-9133-7

[10] Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing. Monographs on Statistics and Applied Probability No. 60* Chapman and Hall. MR1319818

[11] Simonoff, Jeffrey S. (1996). *Smoothing methods in statistics.* Springer-Verlag MR1391963

6. Appendix: a formula for the principal value integral of a ratio where the denominator can pass through zero. Consider the problem of calculating

$$I = \int_A^B dx \frac{a(x)}{b(x)} = \int_A^B dx \frac{a(x)}{\hat{b}(x)} \frac{\hat{b}(x)}{b(x)} \tag{72}$$

where the zeros $\{x_i : 1 \leq i \leq k\}$ (each assumed not to be multiple) of $b(x)$ within $[A, B]$ have been obtained to high precision, and

$$\hat{b}(x) = \prod_j (x - x_j). \tag{73}$$

If all the zeros have been found then the second factor $\hat{b}(x)/b(x)$ in the integral does not pass through zero or a singularity (∞) in the interval $[A, B]$. Using the partial fraction expansion gives

$$\frac{1}{\hat{b}(x)} = \sum_{i=1}^{k} \frac{d_i}{x - x_i}. \tag{74}$$

Multiplying through by $\hat{b}(x)$ gives

$$1 = \sum_{i=1}^{k} d_i \prod_{j=1, j\neq i}^{k} (x - x_j). \tag{75}$$

Setting $x = x_p$, the product is zero if one of the j's is p which is true whenever $i \neq p$, therefore

$$1 = d_p \prod_{j=1, j\neq p}^{k} (x_p - x_j) \tag{76}$$

so

$$d_i = \left(\prod_{j=1, j\neq i}^{k} (x_i - x_j) \right)^{-1} \text{ for } 1 \leq i \leq k. \tag{77}$$

Then substituting (74) into (72) gives

$$I = \sum_{i=1}^{k} \int_A^B dx a(x) \frac{\hat{b}(x)}{b(x)} \frac{d_i}{x - x_i}. \tag{78}$$

To simplify the notation let

$$f(x) = a(x) \frac{\hat{b}(x)}{b(x)} \tag{79}$$

where $f(.)$ can be zero in $[A, B]$ only if $a(.)$ is, but cannot be ∞ there. Then

$$I = \sum_{i=1}^{k} d_i \int_A^B dx \frac{f(x)}{x - x_i} = \sum_{i=1}^{k} d_i \left[\int_A^B dx \frac{f(x) - f(x_i)}{x - x_i} + f(x_i) \ln \left| \frac{B - x_i}{A - x_i} \right| \right] \tag{80}$$

This rearrangement separates the singular term in the integrand so that it can be integrated as a Principal Value integral.

Also $f(x_i)$ is not defined above because its definition becomes $0/0$. It can however be calculated to make $f(.)$ continuous. Using L'hospital's rule

$$\lim_{x \to x_i} \frac{\hat{b}(x)}{b(x)} = \frac{\hat{b}'(x_i)}{b'(x_i)} \tag{81}$$

because $\hat{b}(x_i) = b(x_i) = 0$. The numerator is

$$\hat{b}'(x_i) = \frac{d}{dx} \prod_{j=1}^{k} (x - x_j) \Bigg|_{x=x_i} = \sum_{s=1}^{k} \prod_{j=1, j \neq s}^{k} (x - x_j) \Bigg|_{x=x_i} = d_i^{-1} \tag{82}$$

because the product in the second member is 0 if $i \neq s$ and is d_i^{-1} otherwise. From (79), (81) and (82) it follows that

$$f(x_i) = \frac{a(x_i)}{d_i b'(x_i)}. \tag{83}$$

Hence equation (80), can be written as follows using the results (74), (79) and (83)

$$I = \int_A^B dx \left[\frac{a(x)}{b(x)} - \sum_{i=1}^{k} \frac{a(x_i)}{b'(x_i)(x - x_i)} \right] + \sum_{i=1}^{k} \frac{a(x_i)}{b'(x_i)} \ln \left| \frac{B - x_i}{A - x_i} \right|. \tag{84}$$

107 Science Place
Saskatoon
SK S7N 0X2 Canada

Brook Cottage
The Forge
Ashburnham
Battle
East Sussex
TN33 9PH U.K.