Finite element thin plate splines in density estimation

Markus Hegland[*] Giles Hooker[†] Stephen Roberts[*]

(Received 7 August 2000)

Abstract

The problem of estimating probability density functions differs significantly from functional estimation in which a response variable is present and has for this reason has been dealt with by substantially different methods. We demonstrate here that it is possible to apply

[*] Research School of Information Science and Engineering, Australian National University, Canberra ACT 0200, Australia

[†] School of Mathematical Sciences, Australian National University, Canberra, ACT 0200, Australia

[0]See http://anziamj.austms.org.au/V42/CTAC99/Hegl for this article and ancillary services, © Austral. Mathematical Soc. 2000. Published 27 Nov 2000.

spline-type functionals to the problem of density estimation for large data sets. The resulting estimators may be regarded as kernel methods, but may also be applied to inexact or aggregated data. They can be seen to have moments matching the empirical moments of the data up to the degree of smoothness of the function. Finally, we will show that these functions may be naturally approximated by a finite element method and that doing so will make the method scalable.

Contents

1 Introduction

Nonparametric density estimation is an essential tool in data exploration and the presentation of large data sets. It is used in determining multimodality and skewed-ness in data sets and as a tool for data visualisation. It may also be found in discriminant analysis and in hill-climbing clustering algorithms. The reader is directed to [5] for a comprehensive introduction to common uses and standard techniques in this field. In the context of data mining, we are faced with the need to analyse large and complex data sets, often involving millions of records with possibly thousands of attributes which may be needed in data models. The size and complexity of these data sets requires us to produce methods in which we are assured of scalability with respect to data size.

Historically the oldest, simplest and most widely used estimator is the histogram. If we take a fixed discretisation of the region in which the X_i lie, then for each data point we only need to add 1 to the bin in which it falls to calculate $\hat{f}_n$. This is clearly a scalable routine producing a non-negative estimate and some scaling of $\hat{f}_n$ easily ensures a unit integrand. This is not differentiable (or even continuous), however, and also depends on the origin that we choose for the discretisation, making automated choices of bins difficult [5]. Many estimators have been created based on eliminating the first of these problems. The most sophisticated of these being given originally in [1] on a univariate density as the function minimising $\int (f'(x))^2 dx$ subject to the requirement to match the volume of the histogram on each bin. This has been generalised for higher dimensions and general differential operators in [3] and the methods below may be interpreted as the solution to a related problem.

Suppose we seek to overcome these difficulties with a kernel estimator. For suitable kernels the estimate will be positive, differentiable and with unit integrand, although it is often necessary to forgo positivity in order to achieve an optimal rate of convergence. This has the disadvantage that a naive implementation requires $0(n)$ operations at each data point where the function is to be evaluated. However, a scalable approximation may be obtained using fast Fourier transforms on a uniform grid. See [5] for details.

So far these methods appear somewhat *ad hoc* in the approach that they take to density estimation. They further provide no control over the resulting moments of the estimate that they produce. The aim of this paper is to explore an alternative approach to the estimation of a probability density function that is related to the general spline smoothing problem and hence has some relationship to the histospline problem. In doing so we will show that we are able to match the sample integer moments of arbitrary size. The resulting estimator can be regarded as a kernel estimate. We will also introduce a finite element approximation to this estimate which may be regarded as a generalisation of the orthogonal series estimator (see [5]), which will ensure the scalability of our algorithms. The theoretical framework for these techniques will be developed in §2. We will apply a finite element approximation, in particular a non-conforming approach to the thin plate spline penalty term §3. The results of some numerical experiments will be given in §4 and conclusions drawn in §5.

2 Development of a Spline-Smoothed Density

2.1 A Univariate Example

We will begin by developing a specific model in the univariate case and proceed to a generalised multivariate estimator later. The idea here is to follow a spline approach to functional estimation. The motivation for our estimate is as follows. Let us suppose that we have some initial estimate for the density in L^2 which is based on the data $\{X_i\}_{i=1}^n$, call it f_ϵ. In a technique reminiscent of spline smoothing, we wish to find $\hat{f}_n$ minimising:

$$J_2(u) = \int_\Omega (u(x) - f_\epsilon(x))^2 dx + \lambda \int_\Omega (u''(x))^2 dx$$

on some domain $\Omega \subset \mathrm{R}$. We might regard this as the univariate thin plate spline for continuous data given by f_ϵ. Here the λ will play a similar role to that of the smoothing parameter h in the kernel estimator, but we have made considerably more clear the relationship between smoothness and fidelity that we are looking for. The variational equation associated with this problem is:

$$\int_\Omega s(x)(u(x) - f_\epsilon(x))dx + \lambda \int_\Omega u''(x)s''(x)dx = 0\,, \quad \forall s \in H^2$$

which may be rewritten as:

$$\int_\Omega u(x)s(x)dx + \lambda \int_\Omega u''(x)s''(x)dx = E(s)\,, \quad \forall s \in H^2$$

where $E(s) = \int s(x)f_\epsilon(x)dx$. We will now assume that f_ϵ is close to our data, take the usual estimate:

$$E(s) \approx \frac{1}{n}\sum_{i=1}^{n} s(X_i)$$

and derive the final variational form for our estimator; $\hat{f}_n$ satisfies

$$\int_\Omega \hat{f}_n(x)s(x)dx + \lambda \int_\Omega \hat{f}_n{}''(x)s''(x)dx = \frac{1}{n}\sum_{i=1}^{n} s(X_i) \quad \forall s \in H^2(\Omega). \qquad (1)$$

We will regard this as the defining equation for $\hat{f}_n$; it corresponds exactly to $f_\epsilon = \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}$ where δ_{X_i} is the Dirac delta function centred on X_i. Clearly, such a definition cannot be used in the minimisation of J_2. We have that evaluation is a bounded functional on $H^2(\Omega) \subset C^0(\bar{\Omega})$ and that

$\int u(x)^2 dx + \int u''(x)^2 dx$ is elliptic with respect to the standard norm on H^2. It follows that there is a unique $H^2(\Omega)$ function satisfying (1) (cf. [2]).

It is worth making a few comments about the estimator as defined so far. To begin with, it is clear that the approximation used for $E(s)$ can be modified to take account of inexact or aggregate data. We may be given data already aggregated into a histogram, for example. Alternatively, if we know that there is an error of size ϵ, say, associated with the measurement of the data, we may prefer to use an initial estimate of $f_\epsilon = \frac{1}{2\epsilon n} \sum_{i=1}^n I_{[X_i-\epsilon, X_i+\epsilon]}$ where I_Ω is the indicator function of the set Ω. This would correspond to a naive kernel estimate allowing each X_i to vary within its margin of error, it would also be carried out at the cost of scalability in the finite element approximation we will introduce below.

It can also be seen from (1) that if we let $s = 1$ we immediately have that $\hat{f}_n$ has unit integrand. Equally, $s = x$ shows that $E(\hat{f}_n)$ is equal to the mean of $\{X_i\}_{i=1}^n$. This is quite new as far as non-parametric density estimators are concerned and we will generalise this property to match higher moments as well. Clearly, this correspondence between the empirical moments and those of our estimators will be preserved for other approximations to $E(s)$ in the sense of preserving the moments of f_ϵ.

Suppose we take R as our domain, then we can derive the Euler equations as:

$$u(x) + \lambda u^{(4)}(x) = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}(x)$$

If we have g a Green's function satisfying:

$$g(x) + g^{(4)}(x) = \delta_0(x)$$

then the fundamental solution may be seen to be:

$$\hat{f}_n(x) = \frac{1}{n\lambda^{1/4}} \sum_{i=1}^{n} g\left(\frac{x - X_i}{\lambda^{1/4}}\right)$$

and this approximation is a kernel estimate with kernel

$$g(x) = \frac{1}{\sqrt{2}} e^{\frac{-|x|}{\sqrt{2}}} \left(\cos\left(\frac{|x|}{\sqrt{2}}\right) + \sin\left(\frac{|x|}{\sqrt{2}}\right)\right) .$$

A similar analysis may be performed for the general equations presented below.

2.2 The Generalised Density Spline

In a generalised setting, we use multivariate data and replace $\int u''(x)s''(x)\,dx$ with a bilinear form $C(u,s)$ which we will assume to be continuous on a Hilbert space $H \subset L^2$ on which functional evaluation (at least at our data points) is bounded. Working on d dimensional space, $H^{(d+1)/2}$ for d odd, and $H^{(d/2)+1}$ for d even will be sufficient. Then we take $\hat{f}_n$ to satisfy the Galerkin equations:

$$\int_{\Omega} \hat{f}_n(x)s(x)dx + \lambda C(\hat{f}_n, s) = \frac{1}{n} \sum_{i=1}^{n} s(X_i) \tag{2}$$

Typically we will choose

$$C(u,s) = \int_\Omega Lu(x) \cdot Ls(x)\, dx$$

for some differential operator L. In particular, if we wish to preserve the sample moments up degree r, we can choose L so that

$$Lu \cdot Ls = \sum_{|\alpha|=r+1} \binom{r+1}{\alpha} D^\alpha u D^\alpha s,$$

where the binomial terms ensure rotational invariance of the functional ([6]). In this case our approximation must be in the Hilbert space H^{r+1}.

In the following, we will be interested in a two dimensional distribution over some domain Ω, using a penalty term:

$$C(u,s) = \int_\Omega (u_{x_1,x_1} s_{x_1,x_1} + 2u_{x_1,x_2} s_{x_1,x_2} + u_{x_2,x_2} s_{x_2,x_2}) dx \qquad (3)$$

in which first order moments are preserved.

2.3 Positivity—the Roberts Estimate

Before continuing, we will make a few remarks about positivity. It is clear that in general, the method we have outlined above will not produce a non-negative result. In one particular case, however, it is possible to achieve this.

Theorem 1 *If u satisfies:*

$$\int_\Omega u(x)s(x)dx + \lambda \int_\Omega (\nabla u(x) \cdot \nabla s(x))dx = \hat{E}(s)\,, \quad \forall s \in H^1(\Omega) \tag{4}$$

then u is non-negative.

Here we take $\hat{E}$ to be an $H^1(\Omega)$-bounded functional. In particular, pointwise evaluation is not bounded in H^1 and we are unable to use the standard estimate; taking the integral of s against a histogram generated from the data is a reasonable alternative. This has been labelled the Roberts estimate after the author who recognised this property.

Proof: This is seen by reducing (4) to the Euler equation; u satisfies

$$u(x) - \lambda \Delta u(x) = h(x) \tag{5}$$

where h is the (non-negative) histogram we are integrating s against and has boundary conditions:

$$\frac{\partial u}{\partial n} = 0$$

Letting $\Omega -$ be the set of points where u is negative we have that

$$\frac{\partial u}{\partial n} \geq 0$$

on $\delta\Omega-$. Then we can assert that:

$$\int_{\Omega-} \Delta u(x)dx = \int_{\delta\Omega-} \frac{\partial u}{\partial n}(x)dx \geq 0$$

contradicting the non-negativity of h in (5). ♠

This estimate comes at the penalty of reducing our control over moments to ensuring a unit integrand. Positivity will also not necessarily be maintained in an approximation. Nevertheless, it may be useful to consider.

3 Approximating a Solution

3.1 A Finite Element Discretisation

We have seen that the functions resulting from our estimation techniques result in kernel methods. This is true for more general differential operators and for bounded domains. However, in a more general case, particularly when we wish to place a bound on our domain, the kernels that result will not necessarily be easily computable and may well have to be approximated. An alternative would be to calculate the kernels for the infinite domain and truncate them at the boundary (assuming that $\rho = I_\Omega \rho$), but this would violate the unit integrand requirement. Neither solution offers the prospect of scalability and it is therefore proposed that the fundamental solution to (2) should be approximated using the finite element method.

The usual theorems tell us that (2) will hold on a finite dimensional subspace $V \subset H$. We are then able to derive the usual linear equations on a basis for V. What we notice in doing this is that the data only appears in the load vector, in evaluating $\frac{1}{n}\sum_{i=1}^{n}\phi_j(X_i)$. If we assume that V has dimension $k \ll n$, then the evaluation of the load vector is scalable. The linear system we solve is sparse, so that we obtain a method which may be completed in $O(n+k)$ operations. Making use of the standard conformal finite element method error estimates, this will give us an $0(h^{r+1})$ approximation in $L^2(\Omega)$ to the fundamental solution to the equation in (2) with an rth order differential penalty term.

We note that provided the finite element basis we are using is capable of producing polynomials x^p on the domain Ω, for $0 \le p \le r-1$ (and an rth order B-spline basis will do this nicely) then we can again produce a unit integrand and match the empirical pth moments.

We can, in some sense, regard the finite element approximation as a generalisation of the orthogonal series estimate. If $\{\phi_i\}_{i=1}^{\infty}$ represents an orthogonal system, then substituting this in our finite element calculations, we can regard the method as a choice of tapering parameters for an already-truncated estimator. The method we have, then, allows us to control the interaction between terms in an estimator from a series which is not orthogonal.

3.2 A Non-conforming Method

In the remainder of this paper, we examine the use of a bilinear form as in (3) in our estimate, which would require an approximating subspace to have H^2 basis functions. The proposal here is to adapt the non-conforming method in [4] to this problem and allow us to use only H^1 elements. For $\mathbf{u} = (u_1, u_2) \in H^1(\Omega)^2$ let $P\mathbf{u}$ be a solution to

$$(\nabla v, \nabla P\mathbf{u}) = (\nabla v, \mathbf{u})\,, \quad \forall v \in H^1(\Omega) \tag{6}$$

where $(\cdot,\cdot)$ denotes the usual L^2 inner product. As it stands this problem is not well defined, and we will rectify this by requiring $P\mathbf{u}$ to have zero mean. We will later add in an appropriate constant and use $P\mathbf{u}$ as the final estimate. We will now re-write our variational equations as:

$$(P\mathbf{u}, P\mathbf{s}) + \lambda((u_1, s_1)_{H^1(\Omega)} + (u_2, s_2)_{H^1(\Omega)}) = \frac{1}{n}\sum_{i=1}^{n} P\mathbf{s}(X_i). \tag{7}$$

Theorem 2 *There is a unique solution* $\mathbf{u}$ *to the problem (7).*

Proof: It is clear that the right hand side of (7) is continuous as $P\mathbf{s} \in H^2(\Omega)$ implicitly. Similarly,

$$a(\mathbf{u}, \mathbf{u}) = (P\mathbf{u}, P\mathbf{u}) + \lambda(|u_1|^2_{H^1(\Omega)} + |u_2|^2_{H^1(\Omega)})$$

is also a continuous, positive functional. It remains to show that a is elliptic in the $H^1(\Omega)^2$ norm.

We will begin by denoting $c_1 = \int_\Omega u_1 dx$, $c_2 = \int_\Omega u_2 dx$ and observing that the Poincaré inequality give us that:

$$||\mathbf{u}||^2_{H^1(\Omega)^2} \leq k_1(c_1^2 + c_2^2) + |\mathbf{u}|^2_{H^1(\Omega)^2}$$

We now have that:

$$c_1^2 + c_2^2 \leq k_2 \, [c_1 \;\, c_2] \, X \, [c_1 c_2]^T = k_2 ||c_1 x_1 + c_2 x_2||^2_{L^2(\Omega)}$$

where $[X]_{ij} = \int_\Omega x_i x_j dx_i dx_j$ is a positive definite matrix. Now $P[c_1 \;\, c_2] = c_1 x_1 + c_2 x_2$ and this will also be the case in an H^1 finite element subspace. Hence:

$$||\mathbf{u}||^2_{H^1(\Omega)^2} \leq k_3 \left(||P[c_1 \;\, c_2]||^2_{L^2(\Omega)} + |\mathbf{u}|^2_{H^1(\Omega)^2} \right)$$

Finally we note that:

$$\begin{aligned} ||P[c_1 \;\, c_2]||^2_{L^2(\Omega)} &\leq ||P\mathbf{u}||^2_{L^2(\Omega)} + ||P(\mathbf{u} - [c_1 \;\, c_2])||^2_{L^2(\Omega)} \\ &\leq ||P\mathbf{u}||^2_{L^2(\Omega)} + ||\mathbf{u} - [c_1 \;\, c_2]||^2_{H^1(\Omega)^2} \\ &\leq ||P\mathbf{u}||^2_{L^2(\Omega)} + |\mathbf{u}|^2_{H^1(\Omega)^2}. \end{aligned}$$

Consequently:

$$||\mathbf{u}||^2_{H^1(\Omega)^2} \leq k_4 \left(||P\mathbf{u}||^2_{L^2(\Omega)} + |\mathbf{u}|^2_{H^1(\Omega)^2} \right) \leq k_5 a(\mathbf{u}, \mathbf{u})$$

and the k_i depend only on Ω and λ. It follows that there is a unique solution to this problem and that this result will hold in a non-conforming H^1 approximation. ♠

Constraining this equation to the closed subspace of $H^1(\Omega)^2$ on which $\mathbf{u}$ has zero curl there is again a unique solution. We have that $\mathbf{u} = \nabla(P\mathbf{u})$ and have recovered exactly the original equation (2). We will follow [4] in ignoring this condition.

Applying a finite element approximation on a basis $\{\phi_i\}_{i=1}^k$, the problem (6) may be approximated as a discrete problem in H^1 by:

$$C\mathbf{u}_0^h = B_1\mathbf{u}_1^h + B_2\mathbf{u}_2^h$$

where $[C]_{ij} = (\nabla\phi_i, \nabla\phi_j)$, $[B_k]_{ij} = (\partial\phi_i/\partial x_k, \phi_j)$ and $P_h\mathbf{u} = \sum_{i=1}^k [\mathbf{u}_0^h]_i\phi_i$. Writing

$$\vec{C} = \begin{bmatrix} C & 0 \\ 0 & C \end{bmatrix}, \quad B = [B_1 \; B_2], \quad \mathbf{u}^h = \begin{bmatrix} \mathbf{u}_1^h \\ \mathbf{u}_2^h \end{bmatrix}$$

We can express the discretised (7) in matrix form as:

$$(B^TC^{-1}AC^{-1}B + \lambda\vec{C})\mathbf{u}^h = B^TC^{-1}\mathbf{n}$$

with $[A]_{ij} = (\phi_i, \phi_j)$, $[\mathbf{n}]_j = \sum_{i=1}^n \phi_j(X_i)$. This will now be a $2k$ set of linear equations which will no longer be sparse. It can, however, be solved implicitly by iterative methods requiring only multiplication by sparse matrices.

4 Computational Experiments

We will here make a demonstration of our method on an example problem. We have simulated a tri-model bivariate data set using scaling and translations of the Matlab function “randn” to produce data points. All computations have been done in Matlab on a workstation at ANU.

Implementing a naive Gaussian kernel estimate for this data set, we find that a data set of 1000 points takes a reasonable time to compute whereas the nonconforming estimate and the Roberts estimate both produced results happily with ten times that number. The smoothing parameters in the results below have been chosen visually, with $\lambda = 0.1$ for the spline smoothed densities and $h = 1.2$ for the kernel. We have also used Generalised Cross Validation techniques to automatically chose the smoothing parameter, but have found that for some choices of discretisation (h) the GCV function does not have a well defined minimum. This is due to the fact that h also acts as a smoothing parameter, and so must also be taken into account when choosing the smoothing parameter. We are still continuing our research into a robust choice of α and h simultaneously. As such, for the present study we have chosen the parameters manually for a fixed grid common to all the methods. A visual comparison can be made using the results in Figures 1–5. It is clear that the methods outlined here produce good results and the nonconforming method provides a good fit for the H^2 thin plate density.

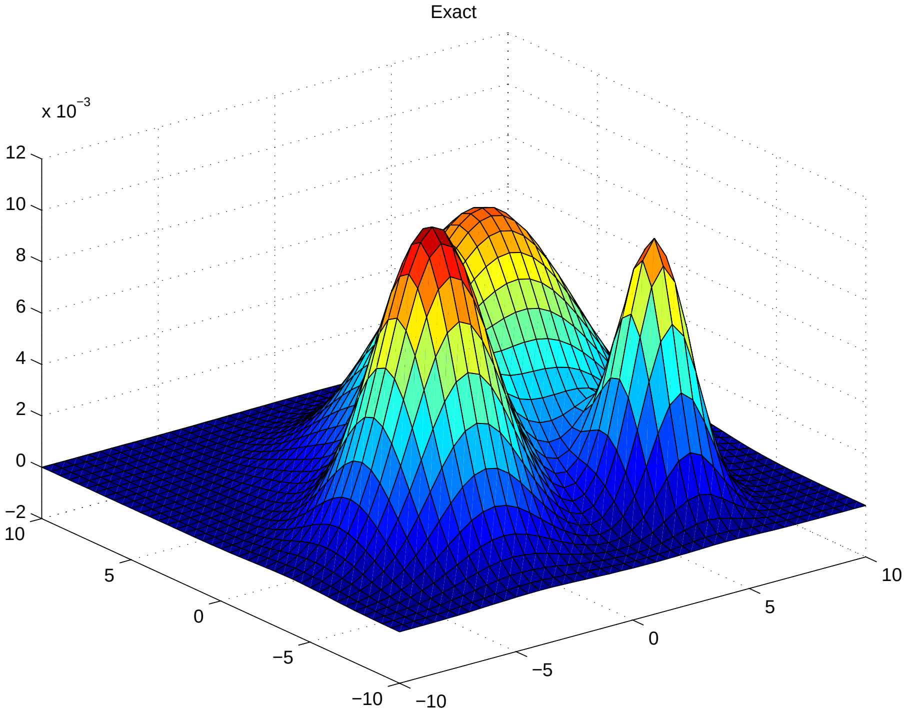

Figure 1: The exact density function

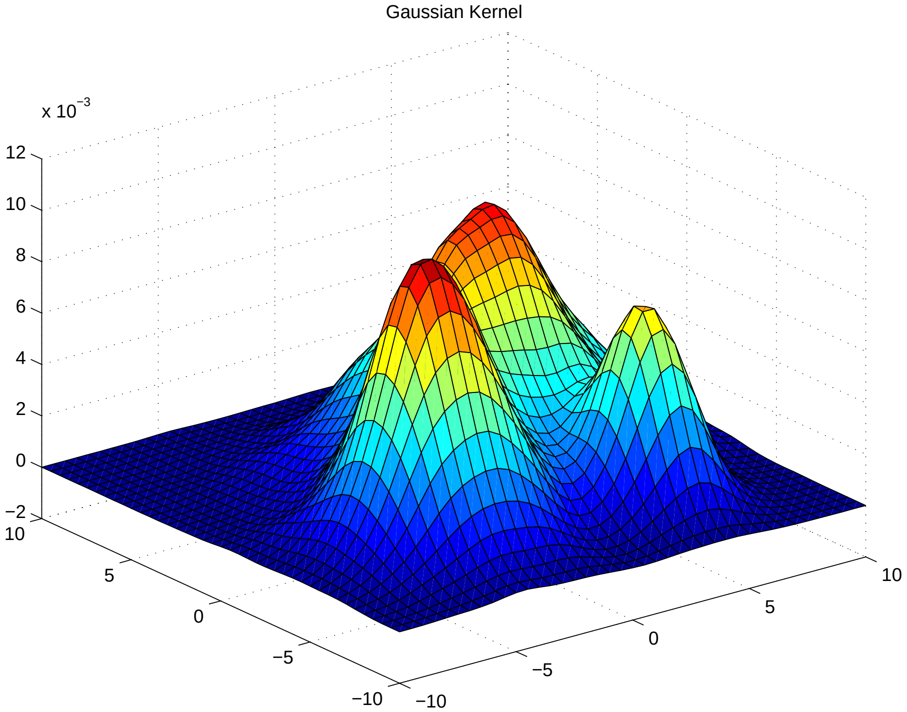

Figure 2: a Gaussian kernel estimate on 1000 points.

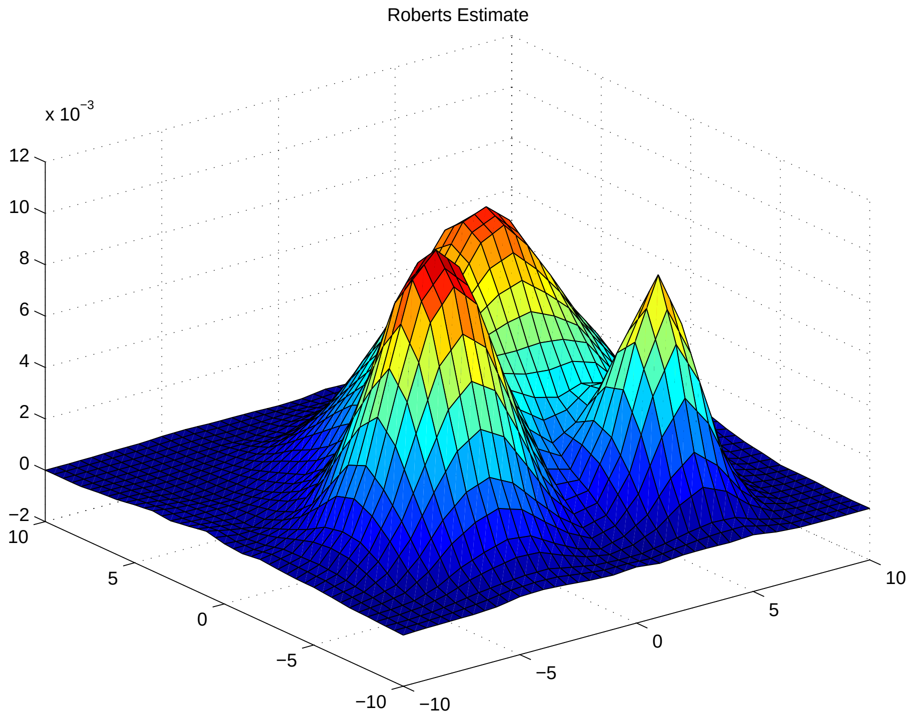

Figure 3: finite element density spline on 10,000 points—Roberts estimate.

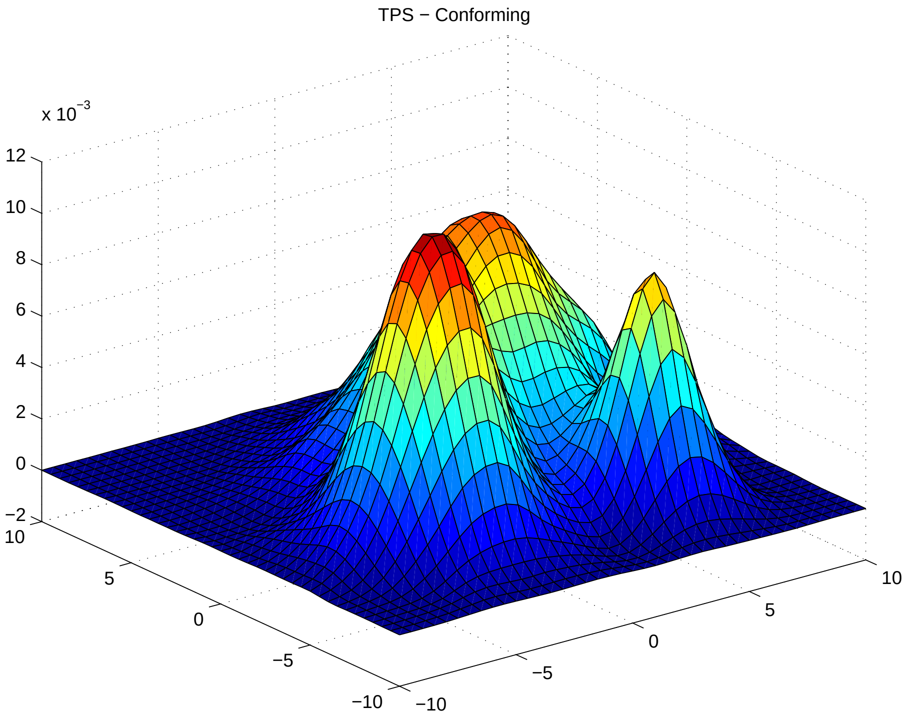

Figure 4: finite element density spline on 10,000 points—TPS - Conforming

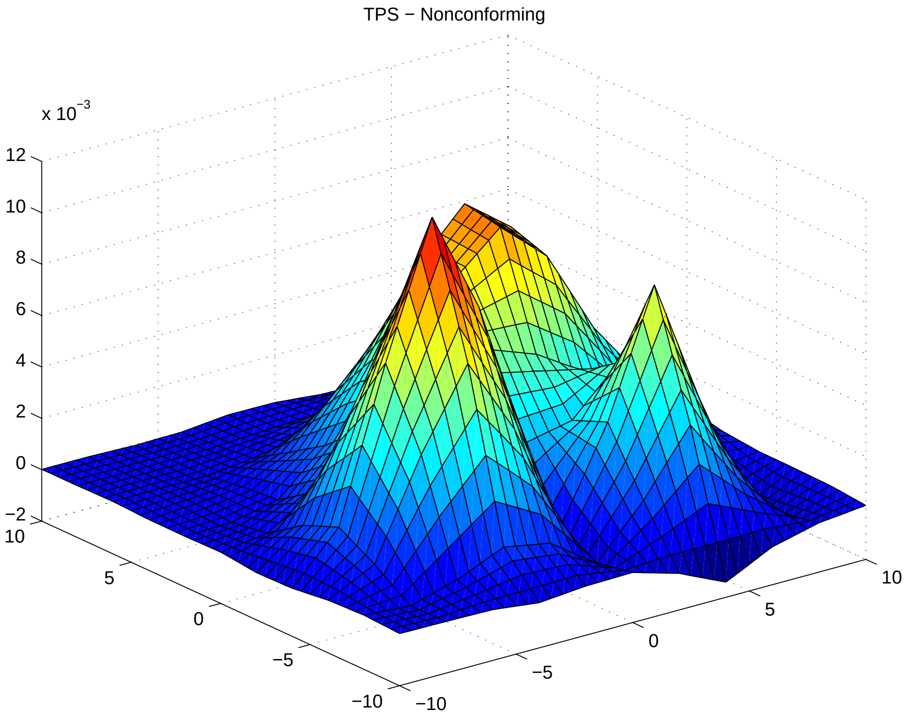

Figure 5: finite element density spline on 10,000 points—TPS - Nonconforming

5 Conclusion

We have here succeeded in producing a density estimator motivated by the use of smoothing splines. We can think of this estimator as an approximation to a kernel method. In employing the finite element method we have created an estimator which is scalable, smooth, global and will fit moments of any given order. We have demonstrated the application of a non-conforming method to the problem with a thin plate spline penalty term and observed that this gives good results and allows us to reduce the complexity of our finite element subspace.

Acknowledgement: This project was supported by the ACSYS CRC.

References

[1] Liliana I. Boneva, David Kendall, and Ivan Stefanov. Spline transformations: Three new diagnostic aids for the statistical data-analyst. *Journal of the Royal Statistical Society*, 33:1–17, 1971. C715

[2] Ph. Ciarlet. *Lectures on the Finite Element Method.* Tata Institute of Fundamental Research, Bombay, 1975. C718

[3] Nira Dyn and Grace Wahba. On the estimation of functions of several variables from aggregated data. *SIAM Journal of Mathematical Analysis*, 13(1):134–152, 1982. C715

[4] M. Hegland, S. Roberts, and I. Altas. Finite element thin plate splines for surface fitting. In John Noye, Michael Teuber, and Andrew Gill, editors, *Computational Techniques and Applications: CTAC97*, pages 289–296, Singapore-New Jersey-London-Hong Kong, 1997. Wold Scientific. C724, C726

[5] B.W. Silverman. *Density Estimation for Statistics and Data Analysis*. Monographs on Statistics and Applied Probability. Chapman and Hall, London-New York, 1986. C714, C714, C715, C715

[6] Grace Wahba. *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathemtics. Society for Industrial and Applied Mathematics, Philadelphia, 1990. C720