Mean conservation for density estimation via diffusion using the finite element method

Conservación de la estimación de densidad vía difusión usando el método de elementos finitos

Keith Yuan Patarroyo Tovar[1,a]

Abstract. We propose boundary conditions for the diffusion equation that maintain the initial mean and the total mass of a discrete data sample in the density estimation process. A complete study of this framework with numerical experiments using the finite element method is presented for the one dimensional diffusion equation, some possible applications of this results are presented as well. We also comment on a similar methodology for the two-dimensional diffusion equation for future applications in two-dimensional domains.

Keywords: mean conservation, diffusion equation, one dimension diffusion, finite element method, perception of security.

Resumen. Proponemos condiciones de frontera para la ecuación de difusión que mantienen la media y la masa total inicial de un conjunto de datos discretos en el proceso de estimación de densidad. Se realizó estudio completo de este esquema con experimentos numéricos usando el método de elementos finitos para la ecuación de difusión en una dimensión, además se resaltaron posibles aplicaciones de estos resultados. También se comentó de implementar una metodología similar para la ecuación de difusión en dos dimensiones para futuras aplicaciones en dominios bidimensionales.

Palabras claves: conservación de media, ecuación de difusión, difusión en una dimensión, método de elementos finitos, percepción de seguridad.

Mathematics Subject Classification: 65M60.

Recibido: enero de 2017 Aceptado: mayo de 2017

1. Introduction

Estimating a density function using a set of initial data points in order to find probability information is a very significant tool in statistics[9]. The method of Kernel Density Estimation (KDE)[11] is now standard in many analysis and applications. Furthermore, this idea has been applied in multiple fields

[1]Departamento de Física, Universidad Nacional de Colombia, Bogotá, Colombia
[a]kypatarroyot@unal.edu.co

(Archaeology [1], Economy [4], etc). The author of this article is particularly interested in constructing Perception of Security (PoS) hotspots using (KDE) methods to analyze real data registered by security experts in Bogotá [5].

Nowadays a wide variety of methods are available to find density functions (KDE) [9],[3]. The method of KDE via difussion is of particular interest for this document; a recent article [2] develops a systematic method for (KDE) using the diffusion equation, also they propose a more general equation to solve some biases for data estimation. However in their analysis, it is only considered the normalization (conservation of mass) of the density function via Neumann boundary conditions, the mean of the sample data is not considered, thus inducing a change of an important initial parameter from the discrete data sample.

In this article, we propose a new set of boundary conditions for the diffusion equation that maintain the initial mean and mass of the the discrete data sample in the density estimation process. A complete study of this framework is performed using the finite element method (FEM) to solve the one-dimensional diffusion equation for different boundary conditions. We show the induced error on the final density when the mean is not conserved. We also show how this one-dimensional model can be used to simulate a (PoS) in a busy avenue of a city. Lastly the new boundary conditions are presented for the two-dimensional diffusion equation for future applications in two dimensional domains.

2. Diffusion equation with different boundary conditions

As it was first noted in [3] and expanded in [2], solving the diffusion equation with a discrete data sample $\{b_n\}_{n=1}^{N}$ as initial condition (2) give an estimate of a continuous probability density function. Then by solving the diffusion equation [8],

$$\begin{cases} \frac{\partial u(x,t)}{\partial t} - \frac{\partial^2 u(x,t)}{\partial x^2} = 0 \quad a < x < b, \quad t > 0, & (1) \\ u(x,0) = \frac{1}{N} \sum_{i=1}^{N} \delta(x - b_i), \qquad x, b_i \in [a,b], & (2) \end{cases}$$

with appropriate boundary conditions and then finding the best t (bandwidth) for the initial data sample one obtains a continuous estimation of the experimental density. In this article we do not consider algorithms for bandwidth selection, we consider only the conservation of the mean. For more information on the bandwidth selection see [2].

This one-dimensional toy problem is nevertheless of interest in applications for constructing (PoS). For instance we can model an avenue as a one dimensional domain where predictions of the most dangerous places in a selected zone can be accomplished.

In the following sections we present the non-conservation of the mean for the Neumann boundary conditions for Problem (1). We also propose new boundary conditions. For the derivations we assume that the functions are sufficiently smooth in order for the theorems of vector analysis to hold. Moreover the following derivations can be done for a more general diffusion equation with a variable diffusion coefficient $k(x)$.

2.1. Neumann boundary conditions

If we consider the Neumann or natural boundary conditions on the Problem (1), we have

$$\frac{\partial u(x,t)}{\partial x}\Big|_a = 0 \quad , \quad \frac{\partial u(x,t)}{\partial x}\Big|_b = 0. \tag{3}$$

As is widely known, the total mass is conserved over time, see Section 2.2, however the mean of the initial condition is, in general, not conserved. Indeed, we have

$$\begin{aligned}\frac{d}{dt}\left(\int_a^b xu(x,t)dx\right) &= \int_a^b x\frac{\partial^2 u(x,t)}{\partial x^2}dx \\ &= \left[x\frac{\partial u(x,t)}{\partial x}\right]_a^b - [u(x,t)]_a^b \\ &= u(a,t) - u(b,t).\end{aligned}$$

Where we used (1), (3) and integration by parts. Hence the mean is generally not conserved, it depends on the values of $u(x,t)$ at the boundary in a time t.

2.2. Boundary conditions that conserve the mean

We propose the following boundary conditions for (1),

$$\frac{\partial u(x,t)}{\partial x}\Big|_a = \frac{\partial u(x,t)}{\partial x}\Big|_b \quad , \quad \frac{u(b)-u(a)}{b-a} = \frac{\partial u(x,t)}{\partial x}\Big|_b. \tag{4}$$

Note that this boundary conditions are non-local, we need to evaluate in both boundary points at the same time. Now we show that both the mean and the mass are conserved over time using this boundary conditions. Consider first the conservation of the total mass. We have,

$$\frac{d}{dt}\left(\int_a^b u(x,t)dx\right) = \int_a^b \frac{\partial^2 u(x,t)}{\partial x^2}dx = \left[\frac{\partial u(x,t)}{\partial x}\right]_a^b = \frac{\partial u(x,t)}{\partial x}\Big|_a - \frac{\partial u(x,t)}{\partial x}\Big|_b = 0.$$

Where we used (1), (4) and integration by parts. This shows that the total

mass is conserved. Consider now the conservation of the mean. We have,

$$\begin{aligned}\frac{d}{dt}\left(\int_a^b xu(x,t)dx\right) &= \int_a^b x\frac{\partial^2 u(x,t)}{\partial x^2}dx \\ &= \left[x\frac{\partial u(x,t)}{\partial x}\right]_a^b - \left[u(x,t)\right]_a^b \\ &= (b-a)\frac{\partial u(x,t)}{\partial x}\Big|_b - u(b,t) + u(a,t) \\ &= 0.\end{aligned}$$

Again (1), (4) and integration by parts were used to obtain the desired result.

This shows that the boundary conditions (4) for problem (1) conserve both mean and mass. Now we proceed to make some numerical simulations using FEM to show the consequences of the application of this boundary conditions in the process of estimation a probability density for a data sample (2).

3. Numerical study of mean conservation

Now the problem (1),(4) is written in a weak formulation [6] in order to apply the finite element method to the problem. Now for all $v(x) \in C^\infty(a,b)$ we have,

$$\int_a^b \frac{\partial u(x,t)}{\partial t}v(x)dx + \int_a^b \frac{\partial u(x,t)}{\partial x}\frac{dv(x)}{dx}dx = (v(b)-v(a))\frac{\partial u(x,t)}{\partial x}\Big|_b. \tag{5}$$

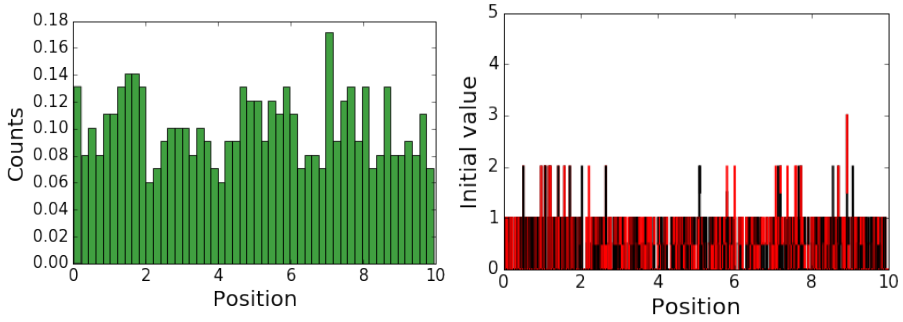
We solve this weak formulation using FEM with low order elements in the interval $[a,b] = [0,10]$, where the number of elements is M. Then Problem (5),(2),(4) yields the problem in the discretised space V^h. Find $u(x,t) \in V^h$, such that for all $v(x) \in V^h$:

$$\begin{cases} \int_a^b \frac{\partial u(x,t)}{\partial t}v(x)dx + \int_a^b \frac{\partial u(x,t)}{\partial x}\frac{dv(x)}{dx}dx = (v(b)-v(a))\frac{\partial u(x,t)}{\partial x}\Big|_b, & (6) \\ u(x,0) = \frac{M}{(b-a)N}\sum_{i=1}^{N}\delta(x-b_i), \quad x, b_i \in [a,b], & (7) \\ \frac{\partial u(x,t)}{\partial x}\Big|_a = \frac{\partial u(x,t)}{\partial x}\Big|_b \quad , \quad \frac{u(b)-u(a)}{b-a} = \frac{\partial u(x,t)}{\partial x}\Big|_b. & (8) \end{cases}$$

Where we represent delta measures by the closest base element of the finite element approximation. Note that (7) contains a normalization factor, since now the elements integral are not one (since they are not delta measures).

Now we use the Galerkin method of mean weighted residuals for the spatial part of the problem choosing low order elements ϕ_i. This formulation can be found in [6]. For our numerical studies we solve the temporal part of the problem (element coefficients) using the implicit-Euler Galerkin Discretization [10], thus the problem is reduced to solve a linear system iteratively for every timestep Δt.

In order to implement the previous formulation numerically, we use `python` to do all the calculations for the simulation. The code is available publicly in [7]. There we start by generating a list of $\{b_n\}_{n=1}^{N=500}$ uniformly distributed points in the interval [0,10]. This points are located in the closest interval of the spatial FEM partition $\{(0+(n-1)/500, n/500)\}_{n=1}^{5000}$. The histogram of this points, Figure 3 can be seen for instance as the number of times a certain criminal act was informed in a zone from the avenue. See Figure 1.



(a) Histogram of the initial discrete data sample using 50 bins.

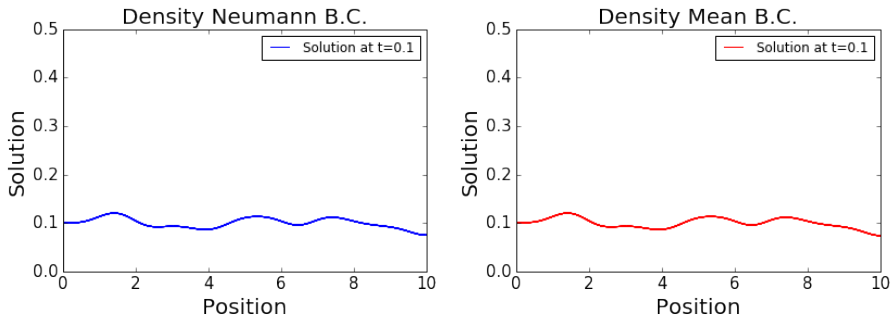
(b) Initial discrete sample date seen as a plot of the FEM initial condition.

Figure 1: Initial discrete data sample to estimate a continuous probability density, its chosen to be uniformly distributed in the interval $[0, 10]$.

If we represent this data as an initial condition (7) we obtain the Figure 1(b). Where we plotted alternatively each consecutive FEM basis function red and black.

Now we solve numerically the problem using the implicit-Euler Galerkin discretization for the problem (6),(7),(8) and we evolve the solution until time $t = 0.1$ using either Neumann boundary conditions, see Figure 2(a) and mean conserving boundary conditions, see Figure 2(b). The solution for the mean conserving boundary condition is positive for this numerical experiment, see Figure 2(b), this fact is currently being explored for future analytical studies.

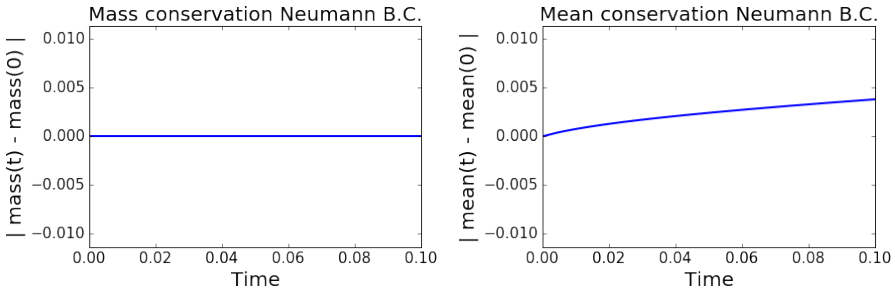
As the Figure 2 shows, the solutions are similar and therefore we can see that for this example the new boundary condition does not generate a noticeable change on the generation of the continuous density distribution. Nevertheless we present the plots of change of mass $\Delta m(t) = m(0) - m(t)$, Figures 3(a), 4(a) and change of mean $\Delta\mu(t) = \mu(0) - \mu(t)$, Figures 3(b), 4(b) for both Neumann and mean conserving boundary conditions.



(a) Numerical solution of the KDE problem using Neumann boundary conditions evolved a time $t = 0.1$.

(b) Numerical solution of the KDE problem using mean conserving boundary conditions evolved a time $t = 0.1$.

Figure 2: Plots of the numerical solution of the problem (6), (7), (8) using both boundary conditions evolved a time $t = 0.1$.

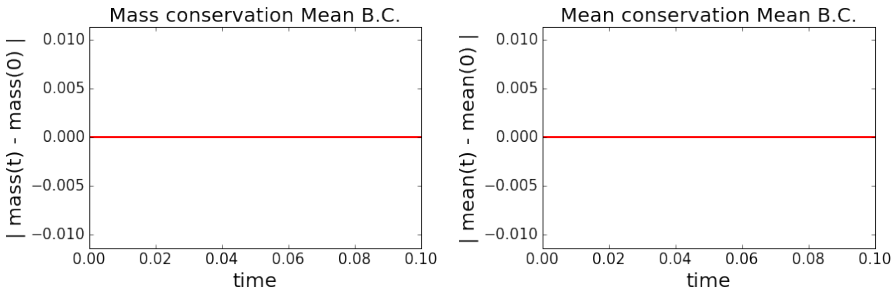


(a) Change of mass Δm for the numerical solution with Neumann boundary conditions.

(b) Change of mean $\Delta\mu$ for the numerical solution with Neumann boundary conditions.

Figure 3: Plots of the evolution of Δm and $\Delta\mu$ for the density estimation with Neumann boundary conditions for $t \in [0, 01]$.

Figures 3(b) and 4(b) present the real difference in the evolution of the density. We effectively see that the mean conserving boundary conditions conserve the mean in the density estimation process. On the other hand if we where to have an initial condition that is biased to one of the boundaries, the differences of the estimated densities by both boundary conditions would differ significantly. However there is no evidence to think that this phenomena occurs in real avenues.



(a) Change of mass Δm for the numerical solution with mean conserving boundary conditions.

(b) Change of mean $\Delta\mu$ for the numerical solution with mean conserving boundary conditions.

Figure 4: Plots of the evolution of Δm and $\Delta\mu$ for the density estimation with mean conserving boundary conditions for $t \in [0, 01]$.

For the numerical experiment presented here we can see that the mean for the Neumann boundary conditions has changed about 0.4% in $t = 0.1$. This change is small, in fact, for an avenue of 10 km, the change in mean would be about 40 m. We conclude that for this numerical experiment for the process of density estimation (when the data has not change to much due to the smoothing process) the Neumann boundary condition provide a very fast (since they are easy to implement) and accurate way to estimate a continuous probability density. Nevertheless the mean of the sample is not preserved exactly, on the other hand, the mean conserving boundary condition, apart from being also easily implementable, is accurate and do preserve the mean of the sample.

4. Two-dimensional densities

We now present the problem for the diffusion equation [8] in two dimensions,

$$\frac{\partial u(\mathbf{x},t)}{\partial t} - \nabla^2 u(\mathbf{x},t) = 0, \qquad \mathbf{x} = (x_1, x_2) \in \Omega \subset \mathbb{R}^2, t > 0. \tag{9}$$

Again we want the conservation of mass and mean in the time evolution of the density. Consider first the conservation of the total mass. We have,

$$\frac{d}{dt}\left(\int_\Omega u(\mathbf{x},t)d\mathbf{x}\right) = \int_\Omega \nabla^2 u(\mathbf{x},t)d\mathbf{x} = \int_{\partial\Omega} \frac{\partial u(\mathbf{x},t)}{\partial \boldsymbol{\nu}} d\sigma,$$

where $\nabla u \cdot \boldsymbol{\nu} = \frac{\partial u(\mathbf{x},t)}{\partial \boldsymbol{\nu}}$, and $\boldsymbol{\nu}$ denotes the outward normal unit vector to $\partial\Omega$. To deduce this relation we used (9), and the first Green identity [8]. Consider

now the conservation of the mean. We have,

$$\frac{d}{dt}\left(\int_\Omega x_i u(\mathbf{x},t)d\mathbf{x}\right) = \int_\Omega x_i \nabla^2 u(\mathbf{x},t)d\mathbf{x} \\ = \int_{\partial\Omega} x_i \frac{\partial u(\mathbf{x},t)}{\partial \boldsymbol{\nu}} d\sigma - \int_\Omega \nabla_i u(\mathbf{x},t)d\mathbf{x},$$

where $\nabla_i u(\mathbf{x},t) = \mathbf{e}_i \cdot \nabla u(\mathbf{x},t)$, assuming Cartesian unit vectors. Again (9) and the first Green's identity were used to obtain the desired result.

Then the conditions that we have to impose on $u(\mathbf{x},t)$ in order to conserve mean and mass are:

$$\int_{\partial\Omega} x_i \frac{\partial u(\mathbf{x},t)}{\partial \boldsymbol{\nu}} d\sigma = \int_\Omega \nabla_i u(\mathbf{x},t)d\mathbf{x}, \quad i = 1,2, \quad \text{and} \quad \int_{\partial\Omega} \frac{\partial u(\mathbf{x},t)}{\partial \boldsymbol{\nu}} d\sigma = 0. \tag{10}$$

The advantage of two dimensional domains is that we are not restricted to impose only two conditions for the boundary(mean and mass conservation). For these domains we can in principle conserve additional higher moments of the density distribution that are meaningful for the particular problem. Applications on two dimensional domains are of special interest for the author since a two dimensional map of the city can generate really robust results in the field of Perception of security(PoS).

5. Conclusions

The proposed mean conserving boundary conditions were shown to effectively maintain the mean of the initial data sample over the continuous density estimation process. This was also confirmed by the numerical simulation of the estimation process where we used a list of uniformly distributed points in the interval [0,10] as an initial condition.

The numerical experiments presented here show that even though Neumann boundary conditions do not conserve the mean over time, they are accurate enough to maintain the mean in a very restricted interval before the oversmoothing of the density estimation process.

We showed the application and some of the consequences of both the idea of (KDE) and the new boundary conditions to avenues in a city. The consequences of implementing the diffusion equation with the proposed boundary conditions in companion of more special initial conditions and in 2D domains remains to be analyzed.

Acknowledgment

I would like to express my gratitude to Juan Galvis, whose guidance was essential for the completion of this manuscript. I also want to thank Francisco

A. Gómez and Zdravko Botev, whose comments were really appreciated for the analysis of the results.
The author thanks the support from project **Hermes 28255**.

References

[1] M. J. Baxter, C. C. Beardah, and S. Westwood, *Sample size and related issues in the analysis of lead isotopedata*, J. Archaeological Science **27** (2000), 973–980.

[2] Z. I. Botev, J. F. Grotowski, and D. P. Kroese, *Kernel density estimation via Diffusion*, Ann. Statist **38** (2010), no. 5, 2916–2957.

[3] P. Chaudhuri and J. S. Marron, *Scale space view of of curve estimation*, Ann. Statist **28** (2000), 408–428, MR1790003.

[4] J. DiNardo, N. M. Fortin, and T. Lemieux, *Labor market institutions and the distribution of wages, 1973 1992: A semiparametric approach*, Econometrica **64** (1996), 1001–1044.

[5] F. Gómez, A. Torres, J. Galvis, J. Camargo, and O. Martínez, *Hotspot mapping for perception of security*, Smart Cities Conference (ISC2), 2016 IEEE International, 2016.

[6] P. Olver, *Introduction to partial differential equations*, Springer-Verlag, 1st Ed., 2014.

[7] K. Patarroyo, *1D Inhomogenious Diffusion Equation solution with FEM*, http://nbviewer.jupyter.org/github/MrKeithPatarroyo/Example/blob/master/FEM-Public/1D%20Heat%20Equation%20FEM-Public.ipynb, 2016, [Online; accessed 23-May-2017].

[8] S. Salsa, *Partial differential equations in action from modelling to theory*, Springer-Verlag, Italia, Milano, 2008.

[9] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London, 1986.

[10] V. Thomee, *Galerkin finite element methods for parabolic problems*, Springer-Verlag, 2nd Ed., 1997.

[11] M. P. Wand and M. C. Jones, *Kernel smoothing*, Chapman and Hall, London, 1995.