PROBABILITY DENSITY FUNCTION ESTIMATION
BY POSITIVE QUARTIC $C^2$-SPLINE FUNCTIONS

Daichi Kitahara and Isao Yamada

Department of Communications and Computer Engineering, Tokyo Institute of Technology, Japan
E-mail: {kitahara, isao}@sp.ce.titech.ac.jp

ABSTRACT

Spline is a continuous function piecewise-defined by polynomials and is widely used for interpolation and smoothing of observed data. In 1994, Heß and Schmidt proposed a positive quartic $C^2$-spline interpolation for estimation of a non-negative and twice continuously differentiable function. In this paper, first we generalize the positive quartic $C^2$-spline interpolation to the positive quartic $C^2$-spline smoothing. Then we propose two estimation methods of a probability density function from its histogram by extending the ideas of the positive quartic $C^2$-spline interpolation and smoothing. Finally numerical experiments show the effectiveness of the proposed methods.

Index Terms— Probability density function estimation, spline function, positive quartic $C^2$-spline smoothing

1. INTRODUCTION

The probability density function (PDF) estimation [1]–[3] is a classical but has been central problem of significant impact to many branches of mathematical sciences and engineering including the Bayesian signal processing [4], pattern recognition [5], [6], quantization [7], filtering [8], data analysis [9], and combustion science [10], etc. Inherently, the PDF estimation requires construction of a positive continuous function from given finite data. The estimation of continuous functions has been a common task in many areas of signal processing where the spline function, a smooth piecewise polynomial, has often been used [11], e.g., in super-resolution of images or videos [12], [13], computer aided design [14], [15], and regression analysis of data [16], [17], due to its flexibility and optimality (see, e.g., Fact 1) in many senses.

However the spline function interpolation has been hardly applicable to the PDF estimation because (i) the data point to be interpolated is not available, and (ii) the positivity of the interpolating spline function is not guaranteed in general. On guaranteeing the positivity of the spline function, Heß and Schmidt showed a way to construct a positive quartic $C^2$-spline function which interpolates given non-negative data samples [18]. These situations suggest that the remaining issue for solid applications of the spline function to the PDF estimation is to establish a more flexible construction of a positive spline function than just interpolating data points.

In this paper, we propose two novel PDF estimations by using positive quartic $C^2$-splines. For this purpose, in Section 2, as a preliminary, we first extend the standard spline smoothing to the positive quartic $C^2$-spline smoothing where the obtained spline function is guaranteed to satisfy an optimality in the sense of a balance between the data fidelity and the smoothness. The proposed PDF estimations are given in Section 3, the first method, for a use in the situation where a relatively rough histogram is available, is realized by modifying the positive quartic $C^2$-spline smoothing shown in Section 2. The second one, for a use in the situation where an almost ideal histogram of which each bin approximates sufficiently corresponding area of the PDF to be estimated is available, is realized by modifying the idea found in [18]. In Section 4, by numerical experiments, we compare the performances of the proposed methods with the kernel density estimation [1], [2] which has been used widely for the PDF estimation. The experiments for a Gaussian mixture show the effectiveness of the proposed methods.

2. PRELIMINARIES

Let $\mathbb{R}$, $\mathbb{R}_+$ and $\mathbb{Z}_+$ denote respectively the set of all real numbers, non-negative real numbers and non-negative integers. A boldface letter denotes a vector or a matrix depending on the situation. For any vector $x \in \mathbb{R}^n$, $|x|_i$ denotes the $i$th component of $x$, and $\|x\| := \sqrt{\sum_{i=1}^n |x|_i^2}$.

2.1. Spline Function

Let $\Delta_n := \{x_i\}_{i=0}^n$ be a grid on an area $\Omega := [x_0, x_n] \subset \mathbb{R}$ s.t. $x_0 < x_1 < \cdots < x_n$ and let $d, \rho \in \mathbb{Z}_+$ s.t. $0 \leq \rho < d$. Define

$S_d^\rho(\Delta_n) := \{ f \in C^\rho(\Omega) | \forall i f = f_i \in P_d \text{ over } [x_i, x_{i+1}] \}$

as the set of all spline functions of degree $d$ and smoothness $\rho$ on $\Delta_n$, where $C^\rho(\Omega)$ stands for the set of all $\rho$-times continuously differentiable functions over the interior of $\Omega$, and $P_d$ denotes the set of all polynomials whose degree is $d$ at most.

Assume that we observe samples of a twice continuously differentiable function $f^* : \Omega \rightarrow \mathbb{R}$ with additive noise $\xi_i \in \mathbb{R}$ on $\Delta_n$, i.e., we observe $\zeta_i := f^*(x_i) + \xi_i$ at $x_i$ ($i = 0, 1, \ldots, n$). In this situation, the spline function is often used as an approximation of $f^*$ because it guarantees the optimality shown in the following fact [19].

Fact 1 (Spline function as a solution of a variational problem) In the set of all twice continuously differentiable functions over the interior of $\Omega$, i.e., $C^2(\Omega)$, there is a unique function $s^* \in C^2(\Omega)$ minimizing

$$\sum_{i=0}^n |\zeta_i - s(x_i)|^2 + \lambda \int_{x_0}^{x_n} |s''(x)|^2 \, dx,$$

and it is a natural cubic spline $s^* \in S_3^2(\Delta_n)$, where $\lambda > 0$ is a parameter controlling the trade-off between the data fidelity and the smoothness of the solution.
2.2. Positive Quartic C²-Spline

The problem of our interest is to approximate a non-negative function \( f^* : \Omega \to \mathbb{R} \) with use of its non-negative data \( \zeta_0 := f^*(x_0) + \epsilon_0 \geq 0 \). In [18], the positive C²-spline interpolation is considered, and it is shown that the lowest degree is \( d = 4 \) for guaranteeing the existence of a spline function \( s \in S^2_d(\Delta_n) \) which satisfies \( s(x) \geq 0 \) for all \( x \in \Omega \) and \( s(x_i) = \zeta_i \) \((i = 0, 1, \ldots, n)\). Therefore in this paper, we use \( S^2_d(\Delta_n) \) as the set of all candidates of estimate of \( f^* \) and introduce the following problem as a natural extension of the standard spline smoothing.

Problem 1 (One-dimensional positive spline smoothing) Find \( s^* \in S^2_d(\Delta_n) \) minimizing
\[
\sum_{i=0}^{n} |s - s(x_i)|^2 + \lambda \int_{x_0}^{x_n} |s''(x)|^2 \, dx
\]
such that
\[
s(x) \geq 0 \quad \text{for all } x \in \Omega.
\]

In order to solve the above problem, we employ the quadratic expression of the integral in the cost function and a sufficient condition [18] for the positivity of \( s \in S^2_d(\Delta_n) \). Define \( z_0 := s(x_0) \), \( p_i := s''(x_i) \) for \( i = 0, 1, \ldots, n \), and define \( h_i := x_{i+1} - x_i \) for \( i = 0, 1, \ldots, n-1 \). Then for ensuring the existence of \( s \in S^2_d(\Delta_n) \), \( z_i, p_i \), and \( P \) to have to satisfy
\[
12z_i + 6h_i^{-1} + h_i^2 P_i = 0 \quad (1)
\]
for \( i = 0, 1, \ldots, n-1 \). The integration in each interval \([x_i, x_{i+1}]\) is expressed as
\[
\int_{x_i}^{x_{i+1}} |s''(x)|^2 \, dx = b_i^T A_i b_i,
\]
where \( b_i := (p_i, p_{i+1}, P_0, P_i)^T \in \mathbb{R}^{4n} \) and
\[
A_i := \begin{pmatrix}
36 & -36 & 3h_i & 3h_i \\
-36 & 36 & -3h_i & -3h_i \\
3h_i & -3h_i & 4h_i^2 & -h_i^2 \\
3h_i & -3h_i & -h_i^2 & 4h_i^2
\end{pmatrix}.
\]

A sufficient condition [18] for the positivity of \( s \in S^2_d(\Delta_n) \) over \( \Omega \) is expressed as
\[
\begin{cases}
z_0 \geq 0, \\
12z_i + h_i p_i + h_i^2 P_i \geq 0 \quad (i = 0, 1, \ldots, n-1), \\
4z_i - h_i p_{i+1} \geq 0 \quad (i = 0, 1, \ldots, n-1), \\
z_{n-1} \geq 0.
\end{cases}
\]

2.3. Solution of Problem 1 under Sufficient Condition

Define \( \zeta, z, p, P \in \mathbb{R}^{n+1}, b \in \mathbb{R}^{2n+2} \) and \( s \in \mathbb{R}^{3n+3} \) as
\[
\zeta := (\zeta_0, \zeta_1, \ldots, \zeta_n)^T
\]
\[
z := (z_0, z_1, \ldots, z_n)^T
\]
\[
p := (p_0, p_1, \ldots, p_n)^T
\]
\[
P := (P_0, P_1, \ldots, P_n)^T
\]
\[
b := (p^T, P^T)^T
\]
\[
s := (z^T, p^T, P^T)^T = (z^T, b^T)^T.
\]
We solve Problem 1 under (2) by reducing to the following convex optimization problem for \( s \).

Problem 1A (Problem 1 under sufficient condition (2)) Find \( s^* \in \mathbb{R}^{3n+3} \) minimizing
\[
\|\zeta - z\|^2 + \lambda b^T A b
\]
such that
\[
E_1 s = 0 \quad \text{and} \quad G s \geq 0,
\]
where \( A \in \mathbb{R}^{(2n+2) \times (2n+2)} \) is a symmetric positive semi-definite matrix constructed by using the components of \( A_k \) and the two constraints \( E_1 s = 0 \) (s.t. \( E_1 \in \mathbb{R}^{n \times (3n+3)} \)) and \( G s \geq 0 \) (s.t. \( G \in \mathbb{R}^{(3n+2) \times (3n+3)} \)) are respectively equivalent to (1) and (2).

Problem 1A can be solved with use of the alternating direction method of multipliers (ADMM) [20]. The ADMM computes \( s^* \) by the following iteration:
\[
\begin{aligned}
s_{k+1} &= \Lambda_1 \left( I - E_1^T (E_1 A E_1^T)^{-1} E_1 \right) s_k + \nu_{k+1} \xi_k + G s_{k+1}, \\
\nu_{k+1} &= P_C (G s_{k+1} + \xi_k)
\end{aligned}
\]
with \( \gamma > 0 \) and any initialization \( s_0 \in \mathbb{R}^{3n+3}, \nu_0 \in \mathbb{R}^{3n+2} \) and \( \xi_0 \in \mathbb{R}^{3n+2} \), where \( I \) is the identity matrix
\[
\Lambda_1 := \left( \begin{pmatrix}
2I & O \\
O & 2\gamma GG^T
\end{pmatrix}
\right)^{-1}
\]
\[
C := \mathbb{R}^{3n+2}, \quad \text{and} \quad P_C \text{ is the metric projection onto } C \text{ and defined as}
\]
\[
[P_C(\nu)_i] := \begin{cases}
[\nu_i] & \text{if } [\nu_i] \geq 0, \\
0 & \text{if } [\nu_i] < 0.
\end{cases}
\]
Finally, we obtain \( s^* \in S^2_d(\Delta_n) \) uniquely from \( s^* \) by computing the coefficients of \( s^* \) (see (3) in Section 3.2).

3. PDF ESTIMATION BY QUARTIC C²-SPLINE

In this section, we propose two estimation methods of a twice continuously differentiable probability density function \( f^* : \mathbb{R} \to \mathbb{R} \) by using the ideas of the positive quartic C²-spline. In this situation, we cannot observe values of \( f^* \) but construct a histogram from observed samples which are generated from \( f^* \) (or we can obtain only the histogram which is published by e.g., the public institution). Hence we reconstruct \( f^* \) from the histogram based on the observed samples.

3.1. Reconstruction of PDF from Its Histogram

Assume that \( \{\eta_k\}_{k=1}^K \) are the observed samples and a grid \( \Delta_n := \{x_i\}_{i=0}^n \) is used as the bins of the histogram s.t. \( x_0 < x_1 < \ldots < x_n, x_0 < \min \{\eta_k\}, \text{ and } x_n > \max \{\eta_k\} \). Let \( K_i \) \((i = 0, 1, \ldots, n-1)\) be the number of \( \eta_k \) which belongs to \( i \)-th bin \([x_i, x_{i+1}])\), i.e., \( x_i \leq \eta_k < x_{i+1} \). Then the histogram roughly approximates samples, and we can expect
\[
\frac{K_i}{K} \approx \frac{1}{x_{i+1} - x_i} \int_{x_i}^{x_{i+1}} f^*(x) \, dx.
\]

Therefore we reconstruct \( f^* \) from the histogram by using the ideas of the positive C²-spline.

We solve Problem 1 under (2) by reducing to the following convex optimization problem for \( s \).

Problem 1A (Problem 1 under sufficient condition (2)) Find \( s^* \in \mathbb{R}^{3n+3} \) minimizing
\[
\|\zeta - z\|^2 + \lambda b^T A b
\]
such that
\[
E_1 s = 0 \quad \text{and} \quad G s \geq 0,
\]
where \( A \in \mathbb{R}^{(2n+2) \times (2n+2)} \) is a symmetric positive semi-definite matrix constructed by using the components of \( A_k \) and the two constraints \( E_1 s = 0 \) (s.t. \( E_1 \in \mathbb{R}^{n \times (3n+3)} \)) and \( G s \geq 0 \) (s.t. \( G \in \mathbb{R}^{(3n+2) \times (3n+3)} \)) are respectively equivalent to (1) and (2).
3.2. Solutions of Problems 2 and 3 under Sufficient Condition

There exists a vector \( c \) and \( s(x) = s'(x_n) = s''(x_0) = s''(x_n) = 0 \), subject to \( s(x) = 0 \) for all \( x \in \Omega \), and

\[
\int_{x_0}^{x_n} s(x) \, dx = 1
\]

\( x = x_0 \) is the optimal solution of Problem 2.

If we can expect

\[
\frac{K_i}{K} = \int_{x_i}^{x_{i+1}} f^*(x) \, dx,
\]

then the formulation in Problem 2 is not enough, and the following formulation is suitable for such cases.

Problems 3 (PDF estimation by positive spline interpolation) Find \( s^* \in S_2(\Delta_n) \) minimizing

\[
\int_{x_0}^{x_n} |s''(x)|^2 \, dx
\]

subject to

\[
s(x) \geq 0 \quad \text{for all} \quad x \in \Omega,
\]

\[
\int_{x_0}^{x_n} s(x) \, dx = 1
\]

and

\[
s(x_0) = s(x_n) = s'(x_0) = s''(x_0) = 0.
\]

where \( \lambda > 0 \).

Moreover by redefining \( \zeta \in \mathbb{R}^n \) as

\[
\zeta := \left( \frac{K_0}{K}, \frac{K_1}{K}, \ldots, \frac{K_{n-1}}{K} \right)^T.
\]

Problem 2 under (2) is expressed as the following convex optimization problem for \( s \).

Problem 2.A (Problem 2 under sufficient condition (2)) Find \( s^* = (z^*, b^*) \) in \( \mathbb{R}^{3n+3} \) minimizing

\[
||\zeta - Hs||^2 + \lambda b^T Ab
\]

subject to

\[
E_2s = 0, \quad Gs \geq 0 \quad \text{and} \quad 1^T Hs = 1,
\]

where \( H := (h_0, h_1, \ldots, h_{n-1})^T \in \mathbb{R}^{n \times (3n+3)} \), \( G \) is defined in Problem 1.A, and the condition \( E_2s = 0 \) (s.t. \( E_2 \in \mathbb{R}^{(s+6) \times (3n+3)} \)) is equivalent to (1) and \( s(x_0) = s(x_n) = s'(x_0) = s''(x_0) = s''(x_n) = 0 \), i.e., \( x_0 = x_n = p_0 = p_n = P_0 = P_n = 0 \), denotes the vector whose all components are 1, and the other condition means \( 1^T Hs = \int_{x_0}^{x_n} s(x) \, dx = 1 \).

The ADMM computes \( s^* \) in the same manner as Section 2.3:

\[
s_{k+1} = \frac{\Lambda_2 J_T^1 (J_1 A_2 J_2^T)^{-1} (0)}{1} + \frac{\Lambda_2 (I - J_T^1 (J_1 A_2 J_2^T)^{-1} J_1 A_2)}{1} \left( 2H^T \zeta + \frac{1}{\gamma} G^T (\nu_k - \xi_k) \right).
\]

\[
\nu_{k+1} = P_C (G s_{k+1} + \xi_k)
\]

\[
\xi_{k+1} = \xi_k + G s_{k+1} - \nu_{k+1}
\]

with \( \gamma > 0 \) and any initialization \( s_0 \in \mathbb{R}^{3n+3} \), \( \nu_0 \in \mathbb{R}^{3n+2} \) and \( \xi_0 \in \mathbb{R}^{3n+2} \), where

\[
J_1 := \left( \frac{E_2}{1^T H} \right) \quad \text{and} \quad J_2 := \frac{2H^T \zeta + \frac{1}{\gamma} G^T (\nu_k - \xi_k)}{1}.
\]

Similarly, Problem 3 under (2) is expressed as follows.

Problem 3.A (Problem 3 under sufficient condition (2)) Find \( s^* = (z^*, b^*) \) in \( \mathbb{R}^{3n+3} \) minimizing

\[
b^T Ab
\]

subject to

\[
E_2s = 0, \quad Gs \geq 0 \quad \text{and} \quad Hs = \zeta.
\]

The ADMM solves Problem 3.A by the following iteration:

\[
s_{k+1} = \Lambda_3 J_T^2 (J_2 A_3 J_3^T)^{-1} (0) \zeta + \frac{1}{\gamma} \Lambda_3 (I - J_T^2 (J_2 A_3 J_3^T)^{-1} J_2 A_3) G^T (\nu_k - \xi_k)
\]

\[
\nu_{k+1} = P_C (G s_{k+1} + \xi_k)
\]

\[
\xi_{k+1} = \xi_k + G s_{k+1} - \nu_{k+1}
\]

with \( \gamma > 0 \) and any initialization \( s_0 \in \mathbb{R}^{3n+3} \), \( \nu_0 \in \mathbb{R}^{3n+2} \) and \( \xi_0 \in \mathbb{R}^{3n+2} \), where

\[
J_2 := \left( \frac{E_2}{H} \right) \quad \text{and} \quad J_3 := \frac{2H^T \zeta + \frac{1}{\gamma} G^T (\nu_k - \xi_k)}{1}.
\]

Remark 1 In some cases, in Problem 3.A, there is no vector which satisfies the constraint. For these cases, there is no solution of Problem 3.A, and the iteration based on the ADMM does not converge. On the other hand, in Problem 2.A, there always exists
In this paper, first we have expanded the idea of the positive quartic form into the convex optimization problems and solved by the ADMM. Then we have proposed two estimation methods of the probability density function from the histogram by the positive quartic -spline interpolation and smoothing. The proposed methods are formulated as the convex optimization problems and solved by the ADMM. Numerical experiments show that the method based on the positive quartic C^2-spline smoothing gives good estimates in the both cases of enough samples and not enough samples.

4. NUMERICAL EXPERIMENTS

Let \( \{\eta_k\}_{k=1}^K \) be observed samples generated from a Gaussian mixture defined as

\[
f^*(x) = \frac{w}{\sqrt{2\pi}\sigma_1^2} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} + \frac{1-w}{\sqrt{2\pi}\sigma_2^2} e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}},
\]

where \((\mu_1, \sigma_1) = (5, 1), (\mu_2, \sigma_2) = (9, 2)\) and \(w = 0.5\). First, we construct a histogram as shown in Fig. 1(a) by using 10000 observed samples \( \{\eta_k\}_{k=1}^{10000} \) and a grid \( \Delta_n := \{x_i\}_{i=0}^{25} \) s.t. \( x_i := 0.5i + 1 \). In Fig. 1(b), \( f^* \) and three estimates \( f_{KDE}, s_{int}^*, s_{simo}^* \) are depicted, where \( f_{KDE} \) is computed by the kernel density estimation [1], [2] using Gaussian kernels as

\[
f_{KDE}(x) := \frac{1}{K} \sum_{k=1}^K \frac{1}{\sqrt{2\pi h^2}} e^{-\frac{(x-\eta_k)^2}{2h^2}},
\]

with \( h = 0.2371 \) and \( s_{int}^* \) and \( s_{simo}^* \) are respectively the solutions of Problems 3.A and 2.A for \( \lambda = 0.015 \). In this case, since there are enough samples, the shape of the histogram has a huge similarity to \( f^* \), and all estimates approximate \( f^* \) with a high degree of accuracy.

Second, we construct a histogram as shown in Fig. 2(a) by using 500 observed samples \( \{\eta_k\}_{k=1}^{500} \) and a grid \( \Delta_n := \{x_i\}_{i=0}^{25} \) s.t. \( x_i := 0.5i + 2 \). In Fig. 2(b), \( f_{KDE} \) and \( s_{int}^* \) are respectively computed by \( h = 0.5048 \) and \( \lambda = 0.015 \). In this case, the shape of the histogram is less similar to \( f^* \), and the values of \( (K_{i+1} - K_i) \) is bigger than the first experiment. Therefore \( s_{int}^* \) does not approximate \( f^* \) and is not smooth. On the other hand, \( s_{simo}^* \) approximates \( f^* \) better than \( f_{KDE} \) even though \( f_{KDE} \) uses the same kernel, Gaussian, as \( f^* \).

5. CONCLUSION

In this paper, first we have expanded the idea of the positive quartic C^2-spline interpolation to the positive quartic C^2-spline smoothing. Then we have proposed two estimation methods of the probability density function from the histogram by the positive quartic C^2-spline interpolation and smoothing. The proposed methods are formulated as the convex optimization problems and solved by the ADMM. Numerical experiments show that the method based on the positive quartic C^2-spline smoothing gives good estimates in the both cases of enough samples and not enough samples.
6. REFERENCES


