Estimation of cumulative distribution function with spline functions

Akhlitdin Nizamitdinov, Aladdin Shamilov

***Abstract*—** The estimation of the cumulative distribution functions (CDF) and probability density functions (PDF) are important in the statistical analysis. In this study we estimated the cumulative distribution functions using following types of spline functions: B-spline, penalized spline (P-spline) and smoothing spline. The data was generated from the mixture normal distributions. We used 15 different mixture normal distributions with broad range of shapes and characteristics. From each model it was generated 1000 samples of sizes n=50, 100 and 200 respectively. We have used the third degree of the spline functions, as it is most used type of spline in recent studies. We compare the estimation accuracy of the three spline estimators with the empirical estimators in terms of their mean squared error (MSE).

***Keywords*—** B-spline, cumulative distribution function, penalized spline, smoothing spline

I. Introduction

REGRESSION analysis is tending to be one of the most popular techniques to estimate data set with unknown function. it is used in various fields, such as, economy, sociology, biology, genetics etc. Most of the problems solving in these fields have a nonlinear effect, i.e. the relationship between variables are nonlinear. These problems could be solved by using parametric nonlinear models, but it gives imprecisely results in estimation. In this case it should be used nonparametric techniques. There are a number of techniques in nonparametric estimation, such as kernel methods, spline functions, etc.

In the past research in estimation of cumulative distribution function the main technique to obtain a smooth nonparametric function is kernel density function. The main idea goes to [1],[2],[3] who made an estimation of univariate independent and identically-distributed data with kernel functions.

Recent studies in kernel estimation of distribution functions were proposed by [4]. This article investigates the estimation of CDF of nonnegative valued random variables using convolution power kernels. Their consistent estimator avoids boundary effects near the origin. Research paper of [5] discusses methods of decreasing the boundary effects, which appear in the estimation of certain functional characteristics of a random variable with bounded support. In the paper of [6] it is discussed utilizing of boundary kernels for distribution function estimation. Also it is investigated the bandwidth selection of kernels as it could be find in [7]. Reference [8] discusses about the smooth simultaneous confidence band construction based on the smooth distribution estimator and the Kolmogorov distribution. Reference [9] studied the asymptotic properties of integrated smooth kernel estimator for multivariate and weakly dependent data.

As far as this paper deals with spline functions, further we give a short literature review on estimation CDF with polynomial and spline functions. There are a few numbers of researches in this area. Reference [10] proposed a smooth monotone polynomial spline (PS) estimator for the cumulative distribution function applied in simulation and real data set. Reference [11] shows estimation of a probability density function using interval aggregated data with spline and kernel estimators. Another research [12] consider the use of cubic monotone control theoretic smoothing splines in estimating the CDF defined on a finite interval. Estimating the probability distribution function with smoothing spline is given in the working paper [13]. She proposed to estimate simulated data from the uniform distribution.

The main objective of this study is to estimate cumulative distribution function by applying different types of spline functions to the values of (X_i) and their function estimation $F(X_i)$, $i = 1, \cdots, n$. We choose regression basis spline model called B-spline. Another type of spline applied in this study is smoothing spline, as it constructs cubic spline basis function and penalty term to control more smoothness in approximation. Penalized spline is a third type of technique that we utilized. It constructs from the B-spline basis function and has a penalty term. The reason of utilizing this method, that it is a combination of two previous one. Obtained results are compared with each other.

The basic idea of regression spline, penalized spline and smoothing splines has described in the following section. Section Simulation study talks about selected functions, proposed methods and results of the analysis.

II. About Spline Functions

The nonparametric regression model has the following form

$$y_i = f(x_i) + \varepsilon_i \quad a < x_1 < \cdots < x_n < b \qquad (1)$$

where $f \in C^2(a, b)$ is an unknown smooth function, y_i, $i = 1, \cdots, n$ are observation values of the response variable y,

A. Nizamitdinov, Anadolu University, Eskisehir, 26470, Turkey (corresponding author, phone:(+90)507-088-8677; e-mail:ahlidin@gmail.com
A. Shamilov, Anadolu University, Eskisehir, 26470, Turkey, (e-mail:asamilov@anadolu.edu.tr)

x_i, $i = 1, \cdots, n$ are observation values of the predictor variable x and ε_i, $i = 1, \cdots, n$ normal distributed random errors with zero mean and common variance σ^2.

The basic aim of the nonparametric regression is to estimate unknown function $f \in C^2(a, b)$ (of all functions f with continuous first and second derivatives) in equation (1). In nonparametric regression, function f is some unknown, smooth function.

Regression spline chooses a basis amounts to choosing some basis functions, which will be treated as completely known: if $b_j(x)$ is a j^{th} such basis function, then f is assumed to have a representation

$$f(x) = \sum_{j=1}^{q} \beta_j b_j(x) \quad (2)$$

for some values of the unknown parameters, β_j

The basic aim of the nonparametric regression is to estimate unknown function $f \in C^2(a, b)$ (of all functions f with continuous first and second derivatives) in model (1). In nonparametric regression, function f is some unknown, smooth function.

B-splines [14],[15] are constructed from polynomial pieces, joined at certain values at the knots τ. Before introducing B-spline basis function let take a short look about Newton's divided difference polynomial.

The n^{th} divided difference of a function f at the points $x_0, \cdots, x_n$, which are assumed to be distinct, is the leading coefficient of the unique polynomial $p_n(x)$ of degree n which satisfies $p_n(x_j) = f(x_j), j = 0, \cdots, n$. The divided difference denoted as $f[x_0, \cdots, x_n]$ or $\Delta_x^n(x_0, \cdots, x_n)f(x)$.

The B-spline $B_{i,k+1}$ of degree k with knots $\tau_i, \cdots, \tau_{i+k+1}$ is defined as

$$B_{i,k+1}(x) = (\tau_{i+k+1} - \tau_i)\Delta_t^{k+1}(\tau_i, \cdots, \tau_{i+k+1})(t - x)_+^k \quad (3)$$

B-spline representation can be expressed as follows:

$$B_{i,k+1}(x) = (\tau_{i+k+1} - \tau_i) \sum_{j=0}^{k+1} \frac{(\tau_{i+j} - x)_+^k}{\prod_{\substack{l=0 \\ l \neq j}}^{k+1} (\tau_{i+j} - \tau_{i+l})} \quad (4)$$

which shows that this function is indeed a spline with $\tau_i, \cdots, \tau_{i+k+1}$ as active knots[15].

Smoothing spline [16] estimate of the function arises as a solution to the following minimization problem: Find $\hat{f} \in C^2(a, b)$ that minimizes the penalized residual sum of squares

$$S(f) = \sum_{i=1}^{n} \{y_i - f(x_i)\}^2 + \lambda \int_a^b \{f''(x)\}^2 dx \quad (5)$$

for some value $\lambda > 0$. The first term in equation denotes the residual sum of the squares and it penalizes the lack of fit. The second term which is weighted by λ denotes the roughness penalty. In other words, it penalizes the curvature of the function f. The λ in (5) is known as the smoothing parameter.

The solution based on smoothing spline for minimum problem in the equation (5) is known as a "natural cubic spline" with knots at $x_1, \cdots, x_n$. From this point of view, a special structured spline interpolation which depends on a chosen value λ becomes a suitable approach of function f in regression model.

Let $f = (f(x_1), \cdots, f(x_n)$ be the vector of values of function f at the knot points $x_1, \cdots, x_n$. The smoothing spline estimate $\hat{f}_\lambda$ of this vector or the fitted values for data $y = (y_1, \cdots, y_n)'$ are given by

$$\hat{f}_\lambda = \begin{bmatrix} \hat{f}_\lambda(x_1) \\ \hat{f}_\lambda(x_2) \\ \vdots \\ \hat{f}_\lambda(x_n) \end{bmatrix} = (S_\lambda)_{n \times n} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad (6)$$

where $\hat{f}_\lambda$ is a natural cubic spline with knots at $x_1, \cdots, x_n$ for a fixed smoothing parameter $\lambda > 0$, and S_λ is a positive-definite smoother matrix which depends on λ and the knot points $x_1, \cdots, x_n$. For general references about smoothing spline, see [16].

The next technique that we used in this study is P-splines [17]. This study makes some significant changes in smoothing spline technique. First, it assumes that $E(y) = \boldsymbol{B}\boldsymbol{a}$ where $B = (B_1(x), B_2(x), \cdots, B_k(x))$ is an $n \times k$ matrix of B-splines and $\boldsymbol{a}$ is the vector of regression coefficients. Secondly, it is supposed that the coefficients of adjacent B-splines satisfy certain smoothness conditions that can be expressed in terms of finite differences of the a_i. Thus, from a least-squares perspective, the coefficients are chosen to minimize

$$S = \sum_{i=1}^{m} \{y_i - \sum_{j=1}^{n} a_j B_j(x_i)\}^2 + \lambda \sum_{j=k+1}^{n} (\Delta^k a_j)^2 \quad (7)$$

For least squares smoothing we have to minimize S in this equation. The system of equations that follows from minimization of S can be written as:

$$\boldsymbol{B}'\boldsymbol{y} = (\boldsymbol{B}'\boldsymbol{B} + \lambda \boldsymbol{D}_k'\boldsymbol{D}_k)\boldsymbol{a} \quad (8)$$

where $\boldsymbol{D}_k$ is a matrix representation of the difference operator Δ^k, and the elements of $\boldsymbol{B}$ are $b_{ij} = B_j(x_i)$.

The problem of choosing the smoothing parameter is one of the main problems in curve estimation. If we use fitting curves by polynomial regression, the choice of the degree of the fitted polynomial is essentially equivalent to the choice of a smoothing parameter. There are a number of different methods to choose smoothing parameter such as, Cross-Validation, Akaike Information For penalized and smoothing splines we used the usual Cross Validation score function. Let $(S_\lambda)_{ii}$ be the i^{th} diagonal element of S_λ. For smoothing splines the usual Cross Validation score function is

$$CV(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{y_i - \hat{f}_\lambda(x_i)}{1 - (S_\lambda)_{ii}} \right\}^2 \quad (9)$$

Here λ is chosen to minimize $CV(\lambda)$.

The selection of the number of knots and their positions are important in approximation problems using spline functions. We used equidistant knots among the data set. The optimal

selection of knot amount could be estimated with Akaike Information Criteria. But in this study we used automatic knots selection with equidistant positions.

III. Simulation Study

This section shows a simulation study that we conducted to evaluate the performance of used techniques. The data was generated from the mixture normal distributions taken from [18]. The selected mixture distributions are shown in Table I.

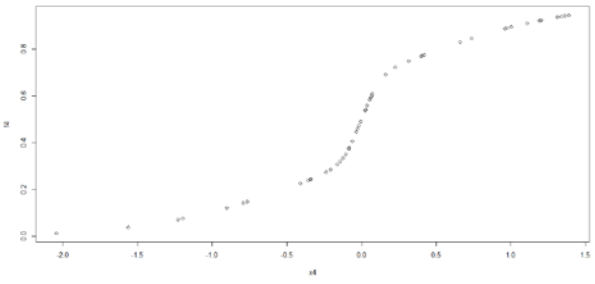
Table I. Mixture normal distributions

Name	Function
Gaussian Density	$N(0,1)$
Skewed Unimodal Density	$F_2(x) = \frac{1}{5}N(0,1) + \frac{1}{5}N\left(\frac{1}{2}, \left(\frac{2}{3}\right)^2\right) + \frac{3}{5}N\left(\frac{13}{12}, \left(\frac{5}{9}\right)^2\right)$
Strongly Skewed Density	$F_3(x) = \sum_{l=0}^{7} \frac{1}{8}N\left(3\left\{\left(\frac{2}{3}\right)^l - 1\right\}, \left(\frac{2}{3}\right)^{2l}\right)$
Kurtotic Unimodal Density	$F_4(x) = \frac{2}{3}N(0,1) + \frac{1}{3}N\left(0, \left(\frac{1}{10}\right)^2\right)$
Outlier Density	$F_5(x) = \frac{1}{10}N(0,1) + \frac{9}{10}N\left(0, \left(\frac{1}{10}\right)^2\right)$
Bimodal Density	$F_6(x) = \frac{1}{2}N\left(-1, \left(\frac{2}{3}\right)^2\right) + \frac{1}{2}N\left(1, \left(\frac{2}{3}\right)^2\right)$
Separated Bimodal Density	$F_7(x) = \frac{1}{2}N\left(-\frac{3}{2}, \left(\frac{1}{2}\right)^2\right) + \frac{1}{2}N\left(\frac{3}{2}, \left(\frac{1}{2}\right)^2\right)$
Asymmetric Bimodal Density	$F_8(x) = \frac{3}{4}N(0,1) + \frac{1}{4}N\left(\frac{3}{2}, \left(\frac{1}{2}\right)^2\right)$
Trimodal Density	$F_9(x) = \frac{9}{20}N\left(-\frac{6}{5}, \left(\frac{3}{5}\right)^2\right) + \frac{9}{20}N\left(\frac{6}{5}, \left(\frac{3}{5}\right)^2\right) + \frac{1}{10}N\left(0, \left(\frac{1}{4}\right)^2\right)$
Claw Density	$F_{10}(x) = \frac{1}{2}N(0,1) + \sum_{l=0}^{4} \frac{1}{10}N\left(\left\{\left(\frac{l}{2}\right) - 1\right\}, \left(\frac{1}{10}\right)^2\right)$
Double Claw Density	$F_{11}(x) = \frac{49}{100}N\left(-1, \left(\frac{2}{3}\right)^2\right) + \frac{49}{100}N\left(1, \left(\frac{2}{3}\right)^2\right) + \sum_{l=0}^{6} \frac{1}{350}N\left(\left(\frac{l-3}{2}\right), \left(\frac{1}{100}\right)^2\right)$
Asymmetric Claw Density	$F_{12}(x) = \frac{1}{2}N(0,1) + \sum_{l=-2}^{2} \frac{2^{1-l}}{31}N\left(l + \frac{1}{2}, \left(\frac{2^{-l}}{10}\right)^2\right)$
Asymmetric Double Claw Density	$F_{13}(x) = \sum_{l=0}^{1} \frac{46}{100}N\left(2l-1, \left(\frac{2}{3}\right)^2\right) + \sum_{l=1}^{3} \frac{1}{300}N\left(\frac{-l}{2}, \left(\frac{1}{100}\right)^2\right) + \sum_{l=1}^{3} \frac{7}{300}N\left(\left(\frac{l}{2}\right), \left(\frac{7}{100}\right)^2\right)$
Smooth Combination Density	$F_{14}(x) = \sum_{l=0}^{5} \frac{2^{5-l}}{63}N\left(\left\{65 - 96\left(\frac{1}{2}\right)^l\right\}/21, \left(\frac{32}{63}\right)^2/2^{2l}\right)$
Discrete Combination Density	$F_{15}(x) = \sum_{l=0}^{2} \frac{2}{7}N\left((12l-15)/7, \left(\frac{2}{7}\right)^2\right) + \sum_{l=8}^{10} \frac{1}{21}N\left(2l/7, \left(\frac{1}{21}\right)^2\right)$

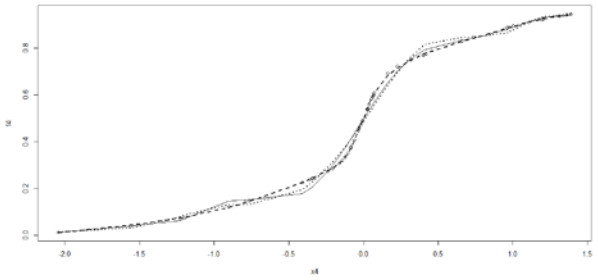
From each of these models, we sampled 1000 samples of sizes n = 50, 100 and 200 respectively. We consider three different type of spline estimators of the distribution function: cubic B- spline, cubic smoothing spline and cubic penalized spline. Unlike the empirical distribution function, all these spline estimators are smooth. We compare the estimation accuracy of the spline estimators with the mean squared error (MSE) performance criteria. For a given $\hat{F}$, the MSE criteria is defined as follow.

$$\mathrm{MSE}\left(\hat{F}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{F}(X_i) - F(X_i)\right)^2 \qquad (10)$$

First, we have generated $n = 50$ samples from each distribution function and estimated with three different spline methods. Fig. 1. demonstrates an example of scatterplot of distribution function values and approximated with spline functions.



a)



b)

Figure 1. a) Scatterplot of distribution function F_4 and b) it's estimation with B-spline (solid line), smoothing spline (dashed line), penalized spline (dotted line)

The estimation results are reported in further tables. Results of estimation of cumulative distribution functions with $n = 50$, $n = 100$ and $n = 200$ are given in Table II, Table III. and Table IV. respectively.

Table II. Estimation results (mean square error values) with $n = 50$.

Functions	B-spline	Smoothing spline	P spline
$F_1(x)$	0.093	0.409	**0.021**
$F_2(x)$	1.257	0.634	**0.439**
$F_3(x)$	60.920	**4.775**	60.035
$F_4(x)$	1000.199	**11.525**	953.043
$F_5(x)$	300.619	**119.285**	271.351
$F_6(x)$	2.042	1.429	**0.483**
$F_7(x)$	8.231	**2.124**	2.550
$F_8(x)$	7.827	**2.290**	3.850
$F_9(x)$	10.054	**3.724**	6.708
$F_{10}(x)$	82.596	**81.488**	87.229
$F_{11}(x)$	2.395	1.849	**0.898**
$F_{12}(x)$	70.871	**22.622**	66.022
$F_{13}(x)$	8.933	7.518	**6.841**
$F_{14}(x)$	80.983	**75.269**	94.602
$F_{15}(x)$	100.468	**60.749**	137.669

From the results shown in Table I. it could be seen that mostly spline with penalty term, i.e. smoothing spline and penalized spline. Actually, outperforming results of spline with penalties is expectable. Because, smoothing term that added to splines with basis functions gives them more smoothness with control of smoothing parameter λ.

Further we give results for simulation with $n = 100$ and $n = 200$ in Table III. and Table IV. respectively.

Table III. Estimation results (mean square error values) with $n = 100$.

Functions	B-spline	Smoothing spline	P-spline
$F_1(x)$	0.241	0.398	**0.053**
$F_2(x)$	3.236	**0.908**	1.111
$F_3(x)$	96.865	**6.711**	80.473
$F_4(x)$	162.759	**9.681**	126.914
$F_5(x)$	152.368	**23.194**	100.935
$F_6(x)$	5.049	**1.082**	1.086
$F_7(x)$	11.532	**1.574**	4.178
$F_8(x)$	15.984	**2.065**	7.611
$F_9(x)$	13.267	**3.318**	9.303
$F_{10}(x)$	101.118	101.170	**90.840**
$F_{11}(x)$	5.043	1.781	**1.436**
$F_{12}(x)$	93.577	**25.558**	80.139
$F_{13}(x)$	13.510	8.080	**7.945**
$F_{14}(x)$	110.587	**75.885**	100.181
$F_{15}(x)$	162.648	**46.625**	200.112

Table IV. Estimation results (mean square error values) with $n = 200$.

Functions	B-spline	Smoothing spline	P-spline
$F_1(x)$	0.564	0.603	**0.134**
$F_2(x)$	7.163	**1.227**	2.374
$F_3(x)$	11.394	**7.5**	10.855
$F_4(x)$	20.923	**8.099**	15.122
$F_5(x)$	38.734	31.446	**28.807**
$F_6(x)$	9.953	**0.794**	2.082
$F_7(x)$	12.619	**1.144**	6.210
$F_8(x)$	27.004	**2.031**	12.751
$F_9(x)$	14.201	**2.671**	12.347
$F_{10}(x)$	111.688	112.701	**106.776**
$F_{11}(x)$	10.04	**1.369**	2.187
$F_{12}(x)$	98.353	**24.346**	92.935
$F_{13}(x)$	19.263	**8.029**	9.126
$F_{14}(x)$	130.408	**76.529**	124.927
$F_{15}(x)$	160.344	**32.489**	285.438

From the given results in tables we can conclude that splines with penalty term show better result than spline function with basis function. Both, smoothing and penalized spline outperform B-spline. As for techniques with penalty term, smoothing spline shows better approximation than penalized spline.

REFERENCES

[1] E.A. Nadaraya, "Some New Estimators for Distribution Functions", in *Theory of Probability and its Applications*,vol. 9, 1964, pp. 497–500.

[2] A. Azzalini, "A Note on the Estimation of a Distribution Function and Quantiles by a Kernel Method", in *Biometrika*, vol. 68, 1981, pp. 326–328.

[3] H. Yamato, "Uniform Convergence of an Estimator of a Distribution Function", in *Bulletin of Mathematical Statistics*, vol. 15, 1973, pp. 69–78.

[4] B. Funke, C. Palmes, "A note on estimating cumulative distribution functions by the use of convolution power kernels", in *Statistics and Probability Letters*, vol. 121, 2016, pp.90–98.

[5] A. Baszczyńska, "Kernel estimation of cumulative distribution function of a random variable with bounded support", in *Statistics In Transition new series*, vol. 17, No. 3, 2016, pp. 541–556.

[6] C. Tenreiro, "Boundary kernels for distribution function estimation", in *Statistical Journal*, vol. 11, No 2, 2013, pp.169–190.

[7] Bruce E. Hansen , Bandwidth Selection for Nonparametric Distribution Estimation, Working paper

[8] J. Wang, F. Cheng, L. Yang, "Smooth simultaneous confidence bands for cumulative distribution functions", in *Journal of Nonparametric Statistics*, vol. 25, No. 2, 2013, pp. 395–407

[9] R. Liu, L. Yang, "Kernel estimation of multivariate cumulative distribution function", in *Journal of Nonparametric Statistics*, vol. 20, No. 8, 2008, pp.661–677
[10] L. Xue, J. Wang, "Distribution function estimation by constrained polynomial spline regression", in *Journal of Nonparametric Statistics*, vol. 22, Issue 4, 2009, pp.443-457.
[11] J.Z. Huang, X. Wang, X. Wu, L. Zhou, "Estimation of a probability density function using interval aggregated data", in *Journal of Statistical Computation and Simulation*, vol.86, pp.3093-3105.
[12] J. K. Charles, S. Sun, C. F. Martin, "Cumulative Distribution Estimation via Control Theoretic Smoothing Splines", in *Three Decades of Progress in Control Sciences*, 2010, pp. 95–104.
[13] E. M. Restle, "Estimating Distribution Functions with Smoothing Splines", working paper
[14] P. Dierckx, *Curve and surface fitting with splines*, Clarendon Press, Oxford, 1993.
[15] C. De Boor, *A Practical Guide to Splines*, New York: Springer, 1978
[16] P.J. Green, B.W. Silverman, *Nonparametric Regression and Generalized Linear Models*, New York: Chapman & Hall, 1994.
[17] P.H.C. Eilers, B.D. Marx "Flexible smoothing using B-splines and penalized likelihood (with comments and rejoinders)". In *Statistical Science*, vol.11, 1996, pp. 89-121.
[18] J.S. Marron, M.P. Wand, "Exact Mean Integrated Squared Error", in *The Annals of Statistics*, vol. 20, 1992, pp. 712–736.