algorithms

MDPI

Article

Nonparametric Estimation of Continuously Parametrized Families of Probability Density Functions—Computational Aspects

Wojciech Rafajłowicz

Department of Computer Engineering, Wroclaw University of Science and Technology, Wyb Wyspianskiego 27, 50 370 Wroclaw, Poland; wojciech.rafajlowicz@pwr.edu.pl

Received: 28 May 2020; Accepted: 4 July 2020; Published: 8 July 2020

Abstract: We consider a rather general problem of nonparametric estimation of an uncountable set of probability density functions (p.d.f.'s) of the form: $f(x; r)$, where r is a non-random real variable and ranges from R_1 to R_2. We put emphasis on the algorithmic aspects of this problem, since they are crucial for exploratory analysis of big data that are needed for the estimation. A specialized learning algorithm, based on the 2D FFT, is proposed and tested on observations that allow for estimate p.d.f.'s of a jet engine temperatures as a function of its rotation speed. We also derive theoretical results concerning the convergence of the estimation procedure that contains hints on selecting parameters of the estimation algorithm.

Keywords: nonparametric estimation; FFT; family of probability density functions

1. Introduction

Consider a family $f(x; r)$ of functions such that for every $r \in [R_1, R_2] \subset \mathbb{R}$ $f(.; r) : \mathbb{R} \to \mathbb{R}$ is a probability density function (p.d.f.) on real line $\mathbb{R}$, while non-random parameter r takes values from a finite interval $[R_1, R_2]$. Assume that we have observations (κ_l, r_l), $l = 1, 2, \ldots d$ at our disposal, where for $r_l \in [R_1, R_2]$ observation κ_l is drawn at random according to p.d.f. $f(x; r_l)$, $x \in \mathbb{R}$. Our aim is to propose, under mild assumptions, a nonparametric estimator of the whole continuum set of p.d.f.'s $\mathcal{F} = \{f(.; r),\ r \in [R_1, R_2]\}$, assuming that the number of data $d \to \infty$ and that r_l's cover $[R_1, R_2]$ more and more densely as $d \to \infty$. Later on, we shall refer to the above stated problem as the $\mathcal{F}$-estimation problem.

In this paper, we concentrate mainly on the algorithmic aspects of the $\mathcal{F}$-estimation problem, since it is computationally demanding. However, we also provide an outline of the proof that the proposed learning algorithm is convergent in the integrated mean squared error (IMSE) sense.

Before providing motivating examples, we briefly indicate similarities, differences, and a generality of this problem among other nonparametric estimation tasks that were considered from the 1950s [1–4]:

1. The $\mathcal{F}$-estimation problem has certain similarities to the nonparametric estimation problem of a bivariate p.d.f. Notice, however, the important difference, namely in our case r is a non-random parameter. In other words, our sub-task is to select r_l's in an appropriate way (or to confine ourselves to such an interval $[R_1, R_2]$ which is covered densely by passive observations of pairs (κ_l, r_l), $l = 1, 2, \ldots d$).
2. One can also notice a formal similarity of our problem and the problem of nonparametric estimation of a non-stationary p.d.f. $f(x; t)$, say that was introduced to the statistical literature by Rutkowski (see [5–7]) and continued in the papers on the concept drift tracking (see [8–10]). There is, however, a fundamental difference between the time t and parameter r. Namely, time

is not reversible and we usually do not have an opportunity to repeat observations at instants preceding present t. On the contrary, in the $\mathcal{F}$-estimation problem, we allow the next observation to be done at $r_{l+1} < r_l$. Furthermore, we allow also for repeated observations for the same value of r.

3. The estimation of several p.d.f.'s was considered in [11], where it was pointed out that it is a computationally demanding problem, because all of these densities should be estimated simultaneously. However, the goal of this paper is quite different than ours. Namely, in [11], the goal was to compare several densities that are not necessarily indexed by the same additional parameter, which does not arise as a parameter of an estimator.
4. Denote by $\hat{f}(.; r)$ nonparametric estimators of $f(., r)$, $r \in [R_1, R_2]$. Having them at our disposal, we immediately obtain nonparametric estimators of a regression function: $\int_{-\infty}^{\infty} x \hat{f}(x; r)\, dx$ with r as the input variable.
5. Similarly, calculating the median of $\hat{f}(.; r)$, we obtain an estimator of the median regression on r. Analogously, estimators of other quantile regression functions can be obtained.
6. When we allow that x and/or r if $f(x; r)$ are vectors, then the $\mathcal{F}$-estimation problem covers also multivariate density and regression estimation problems. In this paper, we do not follow these generalizations, since even for univariate x and r we obtain computationally and data demanding problem. On the other hand, we propose double orthogonal expansion as the base for solving the $\mathcal{F}$-estimation problem. Replacing orthogonal functions of x and r by their multivariate counterparts, we obtain an estimator that formally covers also multivariate cases, but it is still non-trivial to derive a computationally efficient algorithm and to establish its asymptotic properties.

Below, we outline examples of possible applications of the $\mathcal{F}$-estimation:

- The temperature of a jet engine x depends on the rotating speed r of its turbine and on many other non controllable factors. It is relatively easy to collect a large number of observations (κ_l, r_l), $l = 1, 2, \ldots d$ from proper and improper operating conditions of the engine. For diagnostic purposes, it is desirable to estimate the continuum set of p.d.f.s of the temperatures for rotating speeds $r \in [R_1, R_2]$. This is our main motivating example that is discussed in detail in Section 7.1.
- Consider the fuel consumption x of a test exemplar of a new car model. The main impact on x comes from the car speed r, but x also depends on the road surface, the type of tyres and many other factors. It would be of interest for users to know the whole family $\mathcal{F}$ p.d.f.'s in addition to the mean fuel consumption.

  In a similar vain, it would be of interest to estimate $\mathcal{F}$ when x is the braking distance of a car running at speed r.
- Cereal crops x depend on an amount r of a fertilizer applied to a unit area as well as on soil valuation, weather conditions, etc. Estimating $\mathcal{F}$ would allow for selecting r which provides a compromise between a high yield and its robustness against other conditions.

Taking into account a rapidly growing amount of available data and increasing computational power, it would be desirable to extend many other examples of nonparametric regression applications to estimating the whole $\mathcal{F}$.

Our starting point for constructing an estimator for $\mathcal{F}$ is nonparametric estimation of p.d.f.'s by orthogonal series estimators. We refer the reader to the classic results in this field [1,3,12–15]. We shall need also some results on the influence of rounding errors on these estimators (see [16,17]).

The paper is organized as follows: the derivation of the algorithm is presented, a fast method of computation is proposed. Subsequently, tests of synthetic data are preformed and the convergence of the method is shown. Finally, a real world problem regarding jet turbine temperature is presented as well as other possible applications. As an appendix, a detailed proof of convergence is given.

2. Derivation of the Estimation Algorithm

Let us define $(X(r), r)$ as a generic pair, where—for fixed $r \in [R_1, R_2]$—random variable (r.v.) $X(r)$ has p.d.f $f(x;r)$. We use a semicolon ; in order to indicate that r is a non-random variable. For simplicity of the exposition, we assume that $X(r)$ have bounded supports, $[S_1, S_2] \subset \mathbb{R}$ say, which is the same for every $r \in [R_1, R_2]$. Thus, the family $\mathcal{F}$ of p.d.f.'s is defined on $[S_1, S_2] \times [R_1, R_2]$. We additionally assume that f is squared integrable, i.e., $f(.;.) \in L_2([S_1, S_2] \times [R_1, R_2])$.

2.1. Preparations—Orthogonal Expansion

Consider two orthogonal and normalized (orthonormal) sequences $v_k(x), x \in [S_1, S_2]\, k = 1, 2, \ldots$ and $T_j(r), r \in [R_1, R_2]$, $j = 1, 2, \ldots$ that are complete in $L_2(S_1, S_2])$ and $L_2([R_1, R_2])$, respectively. Then, $f(x;r)$ can be represented by the series (convergent in $L_2[S_1, S_2]$) with the following coefficients:

$$a_k(r) = \int_{S_1}^{S_2} v_k(x)\, f(x;r)\, dx, \quad k = 1, 2, \ldots \tag{1}$$

Notice that $a_k(r)$'s can be interpreted as follows:

$$a_k(r) = E_r\left[v_k(X(r))\right], \tag{2}$$

where E_r stands for the expectations with respect to random variable $X(r)$, having p.d.f $f(x; r)$. Furthermore, each $a_k(r)$ can be represented by the series:

$$a_k(r) = \sum_{j=1}^{\infty} \alpha_{kj} T_j(r) \tag{3}$$

with constant coefficients α_{kj}, defined as follows:

$$\alpha_{kj} = \int_{R_1}^{R_2} a_k(r)\, T_j(r) dr, \quad k, j = 1, 2, \ldots \tag{4}$$

Series (3) is convergent in $L_2([R_1, R_2])$. By back substitution, we obtain the following series representation of f:

$$f(x;r) = \sum_{k=1}^{\infty} \sum_{j=1}^{\infty} \alpha_{kj}\, v_k(x)\, T_j(r) \tag{5}$$

The series in (5), convergent in in $L_2([S_1, S_2] \times [R_1, R_2])$, forms a base for constructing estimators for $\mathcal{F}$. They differ in the way of estimating α_{kj}'s and in the way of resolving the so called bias-variance dilemma. The latter can be resolved by appropriate down-weighting α_{kj}'s estimators. In this paper, we confine ourselves to the simplest way of down-weighting, namely, to the truncation of the both sums in a way described later on.

2.2. Intermediate Estimator

The simplest, from the computational point of view, estimator $\hat{f}(x; r)$ for the family $f(x; r)$, $r \in [R_1, R_2]$, we obtain when we additionally assume that the observations of $X(r)$'s are made on an equidistant grid $\rho_1 < \rho_2 < \ldots < \rho_M$ that splits $[R_1, R_2]$ into non-overlapping intervals of the length $\Delta_r > 0$, which cover all $[R_1, R_2]$ in such a way that $R_1 = \rho_1 - \Delta_r/2$ and $R_2 = \rho_M + \Delta_r/2$. In this section, we tentatively impose the restriction that only repeated, but independent and identically distributed (i.i.d.) observations of $X(\rho_m)$, $m = 1, 2, \ldots, M$ are available. In the asymptotic analysis at the end of this paper, we shall assume that M grows to infinity with the number of observations d. Then, also positions of ρ_m's and Δ_r will be changing, but we do not display this fact in the notations, unless necessary.

At each of ρ_m, $m = 1, 2, \ldots, M$, $n_m > 0$ observations $(X_l(\rho_m), \rho_m)$, $l = 1, 2, \ldots, n_m$ are made, keeping the above-mentioned assumptions on mutual independence in force. Additionally, the mutual independence of the following lists of r.v.'s is postulated:

$$\{X_l(\rho_m),\ l = 1, 2, \ldots, n_m\}, \quad m = 1, 2, \ldots, M \tag{6}$$

Then, one can estimate α_{kj}'s by

$$\hat{\alpha}_{kj} = \Delta_r \sum_{m=1}^{M} \hat{a}_k(\rho_m) T_j(\rho_m) \tag{7}$$

where

$$\hat{a}_k(\rho_m) = \frac{1}{n_m} \sum_{l=1}^{n_m} v_k(X_l(\rho_m)) \tag{8}$$

Estimators (7) are justified by (4). Notice that in (7) the simplest quadrature formula is used for approximating the integral $\int_{R_1}^{R_2}$. When an active experiment is applicable, then one can select ρ_m's at nodes of more advanced quadratures with weights, in the same spirit as in [17,18] for nonparametric regression estimators, where the bias reduction was proved. In turn, estimators (8) are motivated by replacing the expectations in (2) by the corresponding empirical means.

Truncating series (5) at K-th and J-th terms and substituting $\hat{\alpha}_{kj}$ instead of α_{kj}, we obtain the following estimator if the family $f(x; r)$, $r \in [R_1, R_2]$

$$\hat{f}(x;\ r) = \sum_{k=1}^{K} \sum_{j=1}^{J} \hat{\alpha}_{kj}\, v_k(x)\, T_j(r). \tag{9}$$

Later on, K and J will depend on the number of observations, but it is not displayed, unless necessary.

Estimator (9) is quite general in the sense that one can select (and mix) different orthonormal bases v_k's and T_j's. In particular, the trigonometric system, Legendre polynomials, Haar system, and other orthogonal vawelets can be applied. The reduction of computational complexity of (9) is possible when, for a given orthogonal system, its discrete and fast counterpart exists. We illustrate this idea in the next section, by selecting v_k's and T_j's as the trigonometric bases, applying discrete Fourier transform (DFT) and the fast Fourier transform (FFT) as its fast implementation.

3. Efficient Learning Algorithm

Our aim in this section is to propose an algorithm for fast calculations of $\hat{\alpha}_{kj}$'s in (9), using the FFT method, which is necessary to learn a proper selection of K and J or a proper selection those of $\hat{\alpha}_{kj}$'s that are essentially different from zero.

3.1. Data Preprocessing

The FFT algorithm operates on data on a grid. Thus, our first step is to attach the set of raw observations $\mathcal{D}_d \stackrel{def}{=} \{(\kappa_l, r_l),\ l = 1, 2, \ldots d\}$ to a grid.

We already have one axis of the grid, namely, points ρ_m, $m = 1, 2, \ldots, M$. Define $\mathcal{B}_m$, $m = 1, 2, \ldots, M$ in the following way. For $j = 1, 2, \ldots, d$ check:

$$x_{mj} \in \mathcal{B}_m \ \text{ iff } \exists (\kappa_l, r_l) \in \mathcal{D}_d : r_l \in [\rho_m - \Delta_r/2,\ \rho_m + \Delta_r/2), x_{mj} = \kappa_l. \tag{10}$$

Notice that the contents of each $\mathcal{B}_m$'s depends on Δ_r, but it is not displayed in the notation. Denote by $\hat{n}_m$ the cardinality of $\mathcal{B}_m$. Clearly, we must have $\sum_{m=1}^{M} \hat{n}_m = d$. Informally, one can consider x_{mj}, $j = 1, 2, \ldots, \hat{n}_m$ in bin $\mathcal{B}_m$ as slightly distorted realizations of r.v.'s in (6).

The second split of the grid goes along the x-axis. Denote by $\chi_1 < \chi_2 < \ldots \chi_N$, $N > 1$ equidistant points such that for $\Delta_x = \chi_2 - \chi_1$ the intervals $[\chi_n - \Delta_x, \chi_n + \Delta_x)$, $n = 1, 2, \ldots, N$ cover the support $[S_1, S_2]$ and $S_1 = \chi_1 - \Delta_x/2$, $S_N = \chi_N + \Delta_x/2$.

Now, we are ready to define the number of observations q_{mn} that are attached to each grid point (χ_n, ρ_m). Namely, q_{mn} is the number of observations x_{mn} that are contained in bin $\mathcal{B}_m$ and simultaneously take values in $[\chi_n - \Delta_x, \chi_n + \Delta_x)$, $n = 1, 2, \ldots, N$, $m = 1, 2, \ldots, M$. Let us define $M \times N$ matrix P with elements:

$$p_{mn} = q_{mn} \Big/ \sum_{l=1}^{N} q_{ml}, \quad n = 1, 2, \ldots, N, \ m = 1, 2, \ldots, M. \tag{11}$$

Clearly, $p_{m.}$ sum up to 1. Notice, however, that—strictly speaking— $p_{m.}$'s are not histograms of r.v.'s $X(\rho_m)$'s, since they are based on the observations contained in bins $\mathcal{B}_m$. Nevertheless, we shall later interpret them as such because – as $\Delta_r \to 0$ – $p_{m.}$'s converge to $f(x; \rho_m)$, assuming that also $\Delta_x \to 0$ in an appropriate way (see [13] for precise statements).

3.2. Fast Calculations and Smoothing

The crucial, mostly time-consuming step is smoothing preprocessed data contained in matrix P. Therefore, it is expedient to apply 2D FFT in order to calculate the DFT of P:

$$G = FFT_{2D}(P; M, N), \tag{12}$$

where the resulting $M \times N$ matrix G has complex elements g_{mn}, $n = 1, 2, \ldots, N$, $m = 1, 2, \ldots, M$ (see, e.g., [19] for the definition of 2D DFT and its implementation as 2D FFT).

Obviously, the inverse transform provides $P = FFT_{2D}^{-1}(G; M, N)$. Thus, in order to smooth P, we have to remove high frequency components from matrix G, retaining only its, appropriately chosen, $K \times J$ sub-matrix $\hat{G}$, $1 < K < M$, $1 < J < N$ and setting to zero other elements of G Instead of setting zeros, one can apply at this stage so-called windowing, e.g., using the Hamming window that provides a mild way of approaching zero.

Remark 1. *The appropriate choice of $K \times J$ sub-matrix $\hat{G}$, with elements $\hat{g}_{kj}$, means that we have to take into account that the complex valued matrix G has the component corresponding to $(0, 0)$ frequency, which is placed as g_{11} (or g_{MN}). Analogously, other components of G, corresponding to low frequencies, are placed at the "corners" of this matrix. Hence, in order to cancel high frequencies, we have to reshape matrix G in order to have $(0, 0)$ frequency in its middle and other low frequencies nearby. Then, to put zeros at the positions corresponding to high frequencies, select sub-matrix $\hat{G}$ and reshape it in the order reverse to that of the reshaping G. This last step is necessary so as to obtain a smoothed version of the P matrix, which would be a $K \times J$ matrix.*

Remark 2. *It is crucial for further considerations to observe that $\hat{\alpha}_{kj}$'s are directly calculable from $\hat{g}_{jk}$'s and their conjugates by adding or subtracting their real and imaginary parts.*

Remark 3. *The choice of K, J or the choice of indexes k, j for which $\hat{alpha}_{kj}$ is left as nonzero, is crucial for proper functioning of the estimation algorithm, since their choice dictates the amount of smoothness of the resulting surface. Below, we discuss possible ways of learning the algorithm to select them properly.*

Although the $\mathcal{F}$ estimation problem differs from the one of estimating bi-variate p.d.f.'s, we may consider the methods elaborated in this field as candidates that might be useful also here.

1. *The cross-validation approach—see [20], where the asymptotic optimality of this method is proved for the trigonometric and the Hermite series density estimators,*

2. *The Kronmal and Tarter rule [21], which—in our case—reads as follows. For* $m = 1, 2, \ldots M$ *and* $l = 1, 2, \ldots N$, *check the following condition:*

$$|g_{kl}|^2 > c/MN, \tag{13}$$

where $c > 0$ *is a preselected constant. According to derivations in [21],* $c = 1$ *, but in the same paper it is advocated to use* $c = 0.5$. *From our simulation experiments, it follows that for moderate* M *and* N $c = 1.5$ *is appropriate, while, for larger* M, N, *even larger constants are better. If for* g_{kl} *condition (13) holds, then leave it in matrix* G *as a nonzero element. Otherwise, set* $g_{kl} = 0$ *in this matrix. Set* $\hat{G} = G$. *Notice that in this case matrix* $\hat{G}$ *is of the same dimensions as* G, *but it has many zeros as its elements.*

Selection of Δ_r and Δ_x (or equivalently M and N) as functions of the number of observations is also very important for the proper estimation. We give some hints on their choice in the section before last, where the asymptotic analysis is discussed.

The performance of Algorithm 1 is illustrated in the next section.

Algorithm 1 Estimation and learning algorithm.

Input: Raw observations (κ_l, r_l), $l = 1, 2, \ldots, d$.
Step 1: Select parameters M, N and $c > 0$ in (13).
Step 2: Perform data preprocessing, as described in Section 3.1, in order to obtain matrix P.
Step 3: Calculate matrix $G = FFT_{2D}(P; M, N)$.
Step 4: Select elements of matrix $\hat{G}$ either by selecting K and J, using cross-validation, or by the Kronmal-Tarter rule.
Step 5: Calculate matrix $\hat{P} = FFT_{2D}^{-1}(P; K, J)$ when in Step 4 the cross-validation is used or as $\hat{P} = FFT_{2D}^{-1}(P; M, N)$ when the Kronmal-Tarter rule is applied.
Output: $\hat{P}$ is the output of the algorithm, if it suffices to have the estimates of $f(x; r)$ on the grid. If one needs the estimates of $f(x; r)$ at intermediate points, then calculate $\hat{\alpha}_{kl}$'s from the corresponding elements of $\hat{G}$ matrix (see Remark 2) and used them in (9).

4. Test on Synthetic Data

The first test can be made using synthetic data. These data are obtained from the family of probability density functions:

$$f_s(x; r) = N_{(\sqrt{r}, 1.)}(x), \tag{14}$$

where $N_{(\mu,\sigma^2)}$ is normal distribution with mean $\mu = r_i$ and variation $\sigma^2 = 1$. The data are generated with 200 points in $r_i = 0., 0.25, ..., 50.$ and 300 random points for each r_i. Those data are binned in order to obtain the matrix of probabilities (11) with size 200×56.

The similarities between two p.d.f.'s can be calculated using distances which are defined in many ways. Here, we would use Hellinger distance and Kullback–Leibler divergence.

For a specific r, the Hellinger distance is defined as follows:

$$H(r)^2 = \frac{1}{2} \int \left(\sqrt{f_s(x; r)} - \sqrt{\hat{f}(x, r)} \right)^2 dx \tag{15}$$

Another integration along r is required in order to obtain distance for all r's:

$$\frac{1}{R_2 - R_1} \int_{r_1}^{r_2} \int_{-\infty}^{\infty} \left(\sqrt{f_s(x; r)} - \sqrt{\hat{f}(x, r)} \right)^2 dx\, dr. \tag{16}$$

The Kullback–Leibler divergence for a specific r can be defined as follows:

$$D_{KL} = \int f_s(x; r)(\log f_s(x; r) - \log \hat{f}(x, r))dx. \tag{17}$$

Again, additional integration along r is required

$$\frac{1}{R_2 - R_1} \int_{r_1}^{r_2} f_s(x;r)(\log f_s(x;r) - \log \hat{f}(x,r))dx\,dr. \tag{18}$$

Due to inherent randomness, the calculations were carried out 100 times. The results are presented in Table 1. Observe that Kullback–Leibler divergence is not symmetric. Its symmetrized version provides almost the same results as in Table 1.

Table 1. Result of distance calculation between synthetic data and reconstruction.

Method	Mean	Standard Deviation
Hellinger	0.00007	3.7×10^{-6}
Kullback–Leibler	0.00014	5.9×10^{-6}

The reconstruction using the presented algorithm can be seen in Figure 1.

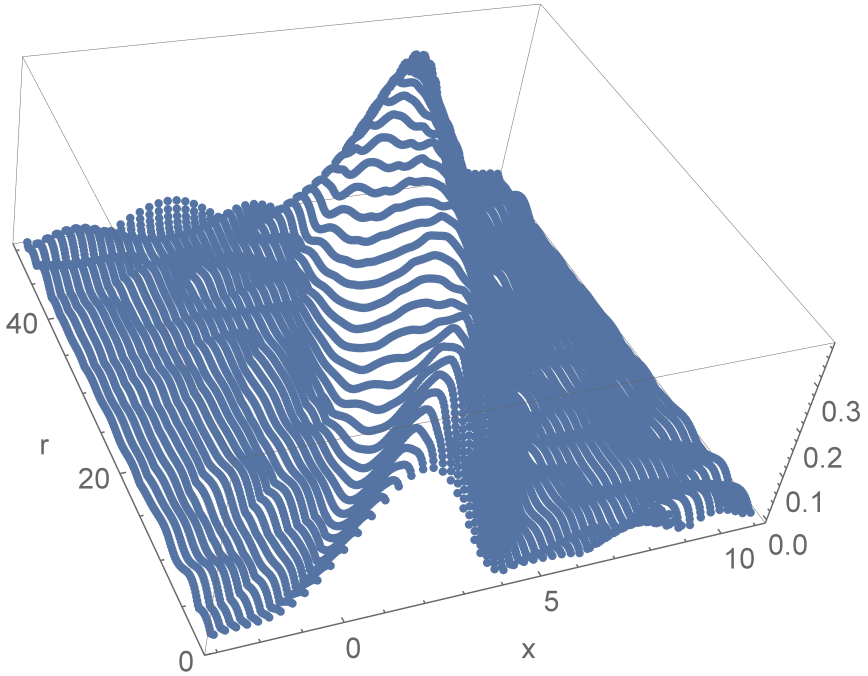

Figure 1. A result for the synthetic problem.

5. Convergence of the Learning Procedure

In this section, we prove the convergence of the learning procedure in a simplified version similar to that described in Section 2.2, but without the discretization with respect to r. Then, we shall discuss additional conditions for the convergence of its discretized version.

Let $X(r), X_1(r), \ldots, X_n$ be independent, identically distributed random variables with parameter r, with p.d.f. $f(., r) \in L^2([R_1, R_2])$, where—for simplicity—we assume that n is the same for each r. Then, f has the representation

$$f(x, r) = \sum_{k=1}^{\infty} a_k(r)\, v_k(x), \tag{19}$$

which is convergent in the $L^2([S_1, S_2])$ norm, where

$$a_k(r) = \int_{S_1}^{S_2} v_k(x)\, f(x, r)\, dx = E_r(v_k(X(r))). \tag{20}$$

Then, we estimate $a_k(r)$'s as follows:

$$\hat{a}_k(r) = \frac{1}{n}\sum_{i=1}^{n} v_k(X_i(r)). \tag{21}$$

Lemma 1. *For every $r \in [R_1, R_2]$ $\hat{a_k}(r)$ is the unbiased estimator of $a_k(r)$.*

Proof. Indeed,

$$E_r(\hat{a_k}(r)) = \frac{1}{n}\sum_{i=1}^{n} E_r[v_k(X_i(r))] = a_k(r). \tag{22}$$

□

As an estimator of $f(., r)$, we take the truncated version of (19):

$$\tilde{f}_n(x,r) = \sum_{k=1}^{K(n)} \hat{a}_k(r)\, v_k(x), \tag{23}$$

where the truncation point K depends on n. It may also depend on r, but, for the asymptotic analysis, we take this simpler version.

The standard way of assessing the distance between f and its estimator $\tilde{f}_n$ is the mean integrated squared error ($MISE$). Notice, however, that, in our case, the MISE additionally depends on r, which is further denoted as $MISE(r)$. Thus, in order to have a global error, we consider the integrated $MISE(r)$ that is defined as follows:

$$IMISE_n = \int_{R_1}^{R_2} MISE_n(r)\, dr. \tag{24}$$

The $MISE(r)$ is defined as follows:

$$MISE_n(r) = E_r \int_{S_1}^{S_2} [f(x, r) - \tilde{f}_n(x, r)]^2 dx, \tag{25}$$

where the expectation w.r.t. $f(., r)$ concerns all $X_i(r)$'s that are present in (23).

Using the orthonormality of v_k's, we obtain:

$$MISE_n(r) = E_r \int_{S_1}^{S_2} [\sum_{k=1}^{K(n)} (a_k(r) - \hat{a}_k(r))^2\, v_k(x) - \sum_{k=K(n)+1}^{\infty} a_k(r)\, v_k(x)]^2 dx \tag{26}$$

Continuing (26), we obtain for each $r \in [R_1, R_2]$:

$$MISE_n(r) = \underbrace{E_r\left[\sum_{k=1}^{K(n)} (a_k(r) - \hat{a}_k(r))^2\right]}_{Var(r, K(n))} + \underbrace{\sum_{k=K(n)+1}^{\infty} a_k(r)^2}_{Bias^2(r, K(n))}. \tag{27}$$

It is known that that for squared integrable f we have: $\sum_{k=1}^{\infty} a_k^2(r) < \infty$. Thus, if $K(n) \to \infty$ when $n \to \infty$, then for every $r \in [R_1, R_2]$ we obtain:

$$Bias^2(r, K(n)) = \sum_{k=K(n)+1}^{\infty} a_k^2(r) \to 0 \text{ when } n \to \infty. \tag{28}$$

This result is not sufficient to prove the convergence of $IMISE_n$ to zero as $n \to \infty$. To this end, we need a stronger result, namely an upper bound on $Bias^2(r\, K(n))$, which is independent of r.

Lemma 2. *Let us assume that* $\frac{\partial}{\partial x} f(x, r)$ *exists and it is a continuous function of* x *in* $[S_1, S_2]$ *for each* $r \in [R_1, R_2]$. *Furthermore, there exists constant* $0 < U < \infty$, *which does not depend on* r, *and such that*

$$\forall\, x \in [S_1, S_2] \;\; \forall r \in [R_1, R_2] \quad \left| \frac{\partial}{\partial x} f(x, r) \right| \leq U. \tag{29}$$

If $K(n) \to \infty$ *when* $n \to \infty$, *then – for* n *sufficiently large we have:*

$$\forall\, r \in [R_1, R_2] \quad Bias^2(r, K(n)) \leq \frac{U^2 (S_2 - S_1)^2}{K(n)}. \tag{30}$$

Proof. It is known that

$$|a_k(r)| \leq k^{-1} \int_{S_1}^{S_2} \left| \frac{\partial}{\partial x} f(x, r) \right| dx \leq k^{-1} U (S_2 - S_1). \tag{31}$$

Thus, $Bias^2(r\, K(n)) \leq U^2 (S_2 - S_1)^2 \sum_{k=K(n)}^{\infty} k^{-2}$, which finishes the proof, since it is known that for sufficiently large $K(n)$ we have $\sum_{k=K(n)}^{\infty} k^{-2} = K^{-1}(n)$. □

For evaluating the variance component, we use Lemma 1:

$$Var(K(n), r) = \sum_{k=1}^{K(n)} E_r[(a_k(r) - \hat{a}_k(r))^2] = \sum_{k=1}^{K(n)} \underbrace{E_r(\hat{a}_k(r) - E_r \hat{a}_k(r))^2}_{Var_r(\hat{a}_k(r))}, \tag{32}$$

where $Var_r(.)$ is the variance of an r.v. having the p.d.f. $F(x, r)$.

Let us assume that the orthonormal and complete system v_k's is uniformly bounded, i.e., there exists p being a non-negative integer and C_p such that

$$\sup_{x \in [S_1, S_2]} |v_k(x)| \leq C_p\, k^p, \quad k = 1, 2, \ldots . \tag{33}$$

Notice that (33) holds for the trigonometric system with $p = 0$

Lemma 3. *If* (33) *holds, then*

$$Var_r(K(n), r) \leq C_p \frac{K^{2p+1}(n)}{n} \tag{34}$$

Proof. $X_i(r)$'s are i.i.d. and (33) holds. Thus,

$$Var_r(\hat{a}_k(r)) = \frac{1}{n^2} \sum_{i=1}^{n} Var_r(v_k(X_i(r))) = \frac{1}{n} Var_r(v_k(X_1(r))) \leq \frac{C_p^2 K^{2p}(n)}{n}. \tag{35}$$

Using this fact in (32) finishes the proof. □

Notice that the bound in (34) does not depend on r.

Theorem 1. *Let the assumptions of Lemmas 2 and 3 hold. If the sequence* $K(n)$ *is selected in such a way that the following conditions hold:*

$$K(n) \to \infty \text{ and } \frac{K^{2p+1}(n)}{n} \to 0 \text{ as an } n \to \infty, \tag{36}$$

then the estimation error $IMISE_n \to 0$ *as* $n \to \infty$.

Proof. By Lemmas 2 and 3, we have uniform (in r) bounds for the variance and for the squared bias, respectively. Thus,

$$MISE_n(r) \leq C_p \frac{K^{2p+1}(n)}{n} + \frac{U^2 (S_2 - S_1)^2}{K(n)}. \tag{37}$$

Hence,

$$IMISE_n \leq (R_2 - R_1) \left[C_p \frac{K^{2p+1}(n)}{n} + \frac{U^2 (S_2 - S_1)^2}{K(n)} \right], \tag{38}$$

which finishes the proof by applying (36). □

Observe that for v_k's being the trigonometric system we have $p = 0$ and the r.h.s. of (38) is, roughly, of the form: $c_1 K(n)/n + c_2/K(n)$, $c1 > 0$, $c_2 > 0$, which is minimized by $K(n) = c_3 \sqrt{n}$ for a certain constant $c_3 > 0$. This implies that $IMISE_n = O(n^{-1/2})$.

6. Proposed Algorithm

Let us define $(x(r), r)$ as a generic par that for fixed r has p.d.f $f(x;r)$. We use semicolon ; in order to indicate that r is a non-random variable.

Observations:

$$(X_i(r_i), r_i), i = 1, 2, \ldots, n \tag{39}$$

we admit that $X_i(r_i)$ and $X_j(r_j)$ are independent random variables even when $r_i = r_j$ for $i \neq j$. In general, r.v.'s in (39) are assumed to be mutually independent.

A family of p.d.f.'s is defined on $[\kappa_1, \kappa_2] \times [R_1, R_2]$ whre $\kappa_1 < \kappa_2$, $R_1 < R_2$ are real numbers. For each $r \in [R_1, R_2]$ $f(x;r)$ is a p.d.f of a random variable $X(r)$.

Consider two orthinormal sequences $v_k(x), x \in [\kappa_1, \kappa_2]$ $k = 1, 2, \ldots$ and $T_j(r), r \in [R_1, R_2]$, $j = 1, 2, \ldots$ that are complete in $L_2([\kappa_1, \kappa_2])$ and $L_2([R_1, R_2])$, respectively. Then, $f(x;r)$ can be represented by the series (convergent in $L_2[\kappa_1, \kappa_2]$

$$a_k(r) = \int_{\kappa_1}^{\kappa_2} v_k(x) f(x;r) dx, \; k = 1, 2, \ldots \tag{40}$$

Notice that $a_k(r)$'s can be interpreted as

$$a_k(r) = E_r v_k(X(r)), \tag{41}$$

where E_r stands for the expectations with respect to random variable $X(r)$ with p.d.f $f(x;r)$. Furthermore, each $a_k(r)$ can be represented by the series:

$$a_k(r) = \sum_{j=1}^{\infty} \alpha_{kj} T_j(r) \tag{42}$$

that is convergent in $L_2([R_1, R_2])$ for constant coefficents α_{kj} defined as

$$\alpha_{kj} = \int_{R_1}^{R_2} a_k(r) T_j(r) dr, \; k, j = 1, 2, \ldots \tag{43}$$

By the back substitution, we obtain the following series representation (in $L_2([\kappa_1, \kappa_2] \times [R_1, R_2])$)

$$f(x;r) = \sum_{k=1}^{\infty} \sum_{j=1}^{\infty} \alpha_{kj} v_k(x) T_j(r) \tag{44}$$

The simplest from the computational point of view, estimator $\hat{f}(x;r)$or the family $f(x;r)$, $r \in [R_1, R_2]$ we obtain, when we additionally assume that the observations of $X(r)$'s are made on an equidistant grid $\rho_1 < \rho_2 < \ldots < \rho_M$ that splits $[R_1, R_2]$ into nonoverlapping intervals of the length $\Delta > 0$.

At each of ρ_m, $m = 1, 2, \ldots M$, n_m observations $(X_k(\rho_m), \rho_m)$, $l = 1, 2, \ldots, n_m$ are made, keeping the above-mentioned assumptions on mutual independence in force.

Then, one can estimate α_{kj}'s by

$$\hat{\alpha}_{kj} = \Delta \sum_{m=1}^{M} \hat{a}_k(\rho_m) T_j(\rho_m) \tag{45}$$

where

$$\hat{a}_k(\rho_m) = \frac{1}{n_m} \sum_{l=1}^{n_m} v_k(x_l(\rho_m)) \tag{46}$$

Estimators (46) are justified by (42). Notice that in (46) the simplest quadrature formula is used for approximating the integral $\int_{R_1}^{R_2}$. When an active experiment is applicable, then one can select ρ_m's at nodes of more advanced quadratures with weights, in the same spirit as in [17] for nonparameteric regression estimators.

Truncating series (44) at K-th and J-t terms and substituting $\hat{\alpha}_{kj}$ instead of α_{kj}, we obtain the following estimator if the family $f(x;r)$, $r \in [R_1, R_2]$

$$\hat{f}(x;r) = \sum_{k=1}^{K} \sum_{j=1}^{J} \hat{\alpha}_{kj} v_k(x) T_j(r). \tag{47}$$

If as $T_j(r)$ a trigonometric series is used on $[R_1, R_2]$, then $\hat{\alpha}_{kj}$ can be calculated using FFT. In addition, v_k can be trigonometric too.

7. Estimating Jet Engine Temperature Distributions

7.1. Subject Overview

A jet turbine is a well-known engine known for at least 90 years The typical application is that of an aircraft power-plant, but also in some ground applications. The main examples are firefighting and snow removal. In recent years, some companies have started to develop engines optimized for thermal power rather than thrust.

In a very simplistic view (see Figure 2), an already running turbine has one input parameter that is fuel flow. As outputs, we have its rotational speed (given in rpm—r) and temperature (in °C—T). In ideal conditions, there should exist a simple relationship between T and r. Since many factors vary and not all of them can be directly measured then we can think about this relationship as probability, which puts us in the framework of the problem stated in the introduction.

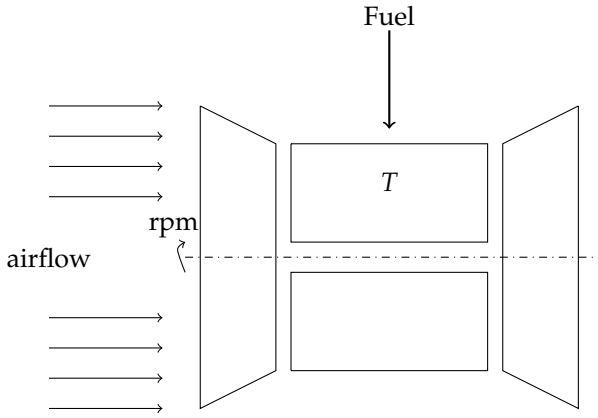

Figure 2. Simplified schematics of turbine—parts important for current investigations.

7.2. Data Preparation

The orginal process data from the server have a form of JSON file of a database table. In this table, we are interested only in two columns, namely the turbine rotation speed (in `rpm`) and the turbine temperatue. Since the resulting temperature is dependent on many other factors, not all of them measured or even known, we treat the value as a random variable X and the rotation speed as known (so not random) r. The amount of relevant data are 71,268.

These two columns have to be changed into frequencies for each r_i. Groups are formed on a rectangular grid. An example of such a grid can be seen in Figure 3. The resulting number of samples in each box of the grid is shown in Figure 4. They should be converted into frequencies by simple divison. In order not to obscure the image, a 32×32 grid was used.

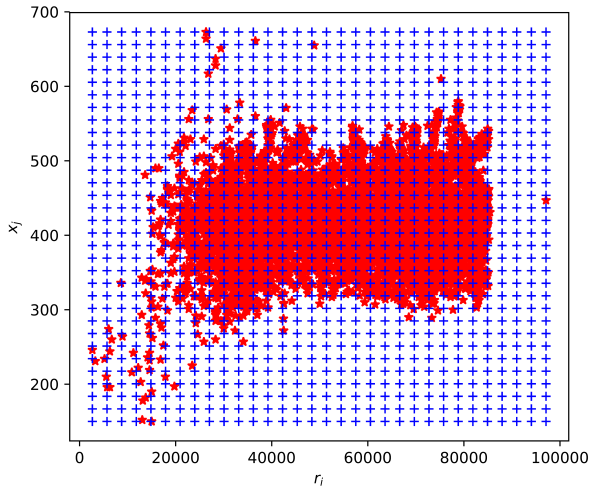

Figure 3. Data and 32×32 grid for frequencies.

As a result, we obtain a matrix containing the observations near points on the grid $\rho_1, \ldots, \rho_m$

$$\begin{bmatrix} f(X_{11}, \rho_1) & f(X_{21}, \rho_1) & \ldots & f(X_{n1}, \rho_1) \\ f(X_{12}, \rho_2) & f(X_{22}, \rho_1) & \ldots & f(X_{n2}, \rho_2) \\ \ldots & \ldots & \ldots & \ldots \\ f(X_{1m}, \rho_m) & f(X_{2m}, \rho_m) & \ldots & f(X_{nm}, \rho_m) \end{bmatrix} \tag{48}$$

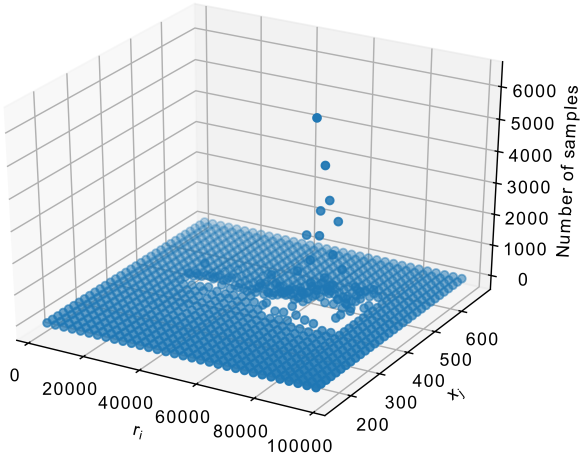

Figure 4. Number of samples in the grid (reduced to 32 × 32 for display's sake).

7.3. The Estimation Process

From the matrix obtained in the previous section, a two-dimensional Discrete Fourier Transform is calculated using the FFT algorithm. The result is a matrix $F_{m\,n}$ of equal size but containing complex values. In literature regarding the Fourier transform, many properties can be found. A good example is book [22]. We use symmetry and antisymmetry of the resulting matrix. General inversion of the Fourier transform is defined in the following way:

$$f(i,j) = \frac{1}{\sqrt{m \cdot n}} \sum_{k=0}^{m} \sum_{l=0}^{n} F_{k,l} \exp(\mathbf{j} \frac{k \cdot i}{m} \cdot \frac{l \cdot j}{n}) \tag{49}$$

where $\mathbf{j}^2 = -1$.

Equation (49) is defined only for the original points of the matrix (48). The continuous surface can be obtained by changing the therms $\frac{i}{m} \in [0,1]$, $\frac{j}{n} \in [0,1]$ into $\frac{r}{\max r_i} \in [0,1]$, $\frac{T}{\max T_i} \in [0,1]$. We cannot guarantee that between grid points the result would be real. The sensible solution is to use the absolute value of a possibly complex number.

We can ask the question of why use the FFT, if we reconstruct the same data again. The reason is smoothing the result. The reconstruction can use only a selected part of the spectrum—obviously lower frequencies. As mentioned before, they reside in the center of the matrix:

$$f(r,T) = \left| \frac{1}{\sqrt{m \cdot n}} \sum_{k=K}^{m/2+K} \gamma_k \sum_{l=L}^{n/2+L} F_{k,l} \exp(\mathbf{j} \frac{k \cdot r}{r_{max}} \cdot \frac{l \cdot T}{T_{max}}) \right|, \tag{50}$$

where γ_k is the correction factor necessary for compensating for the omitted data. Its values should be selected so the reconstruction result is still p.d.f. and integrates to 1 . Obviously, the size of part of the FFT matrix taken is $4 \cdot K \cdot L$, the careful selection of K, L is crucial. When those numbers are too small (see Figure 5), the result loses any resemblance to the original data (Figure 4). On the other hand, when the numbers are too large, this gives a detailed reconstruction (Figure 6) with all unnecessary details.

The acceptable reconstruction in Figure 7 is made with $K = 16$ and $L = 4$. It is obvious that smaller size means faster calculations.

The exact values can be obtained experimentally (as in this example) or by using some method like the Kronmal–Tarter rule.

Another heuristic approach is to reduce the abount of the entry points (probabilities) $f(X, \rho)$ by removing perimeter data. The result of such an approach can be seen in Figure 8. The removal of peripherial data results in reduction of over-fitting.

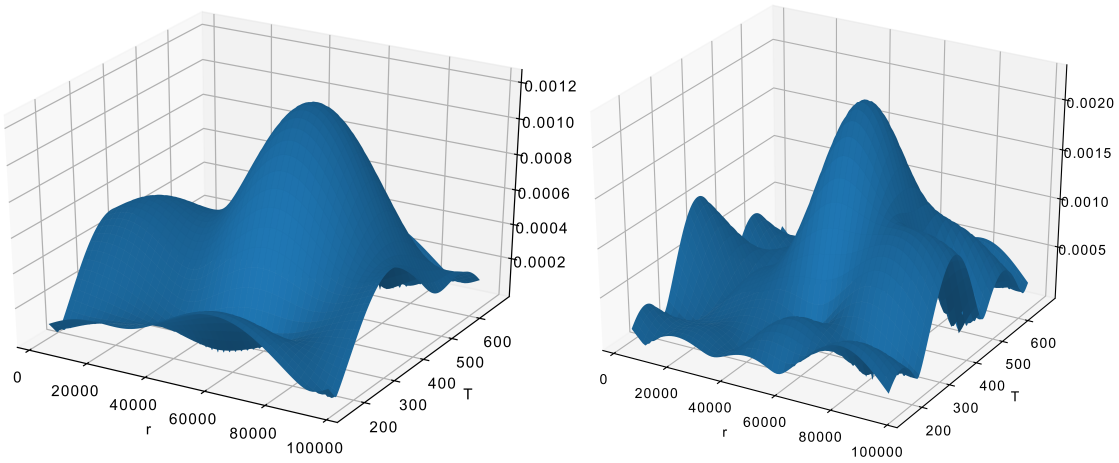

Figure 5. Insufficient amount of Fourier-transformed data, the result is too smooth (2×2 in the **left**, 4×4 in the **right**).

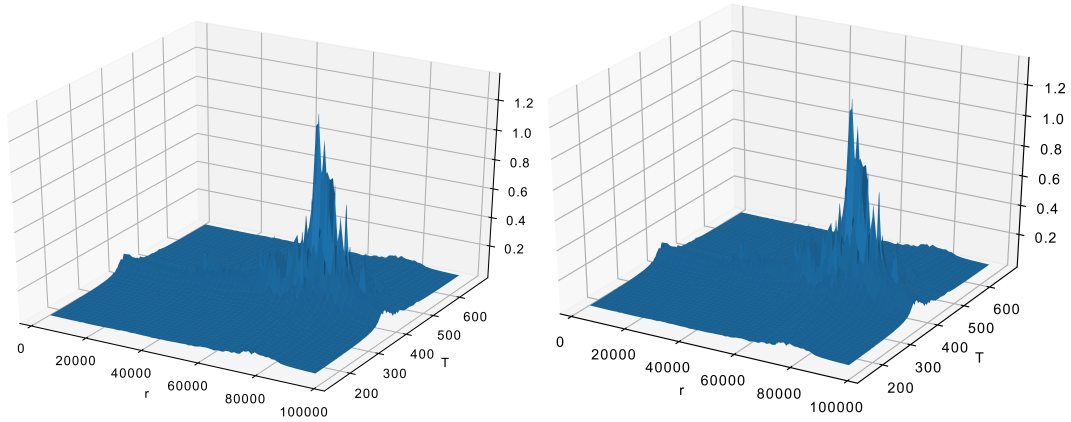

Figure 6. Use of too much Fourier-transformed data (64×64 in the **left**) and full reconstruction (**right**).

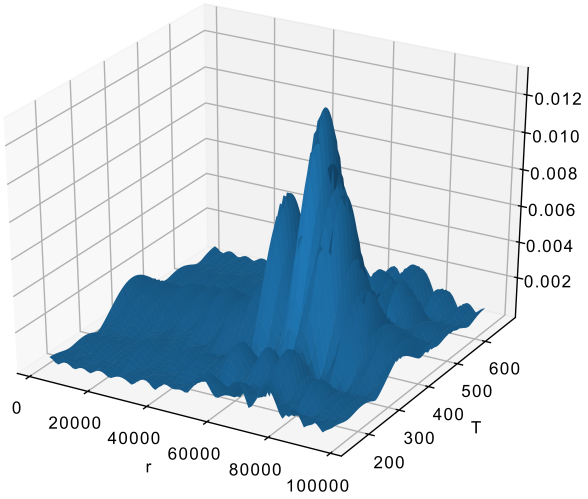

Figure 7. Good selection of data for smoothing, 16 in r, 4 in T.

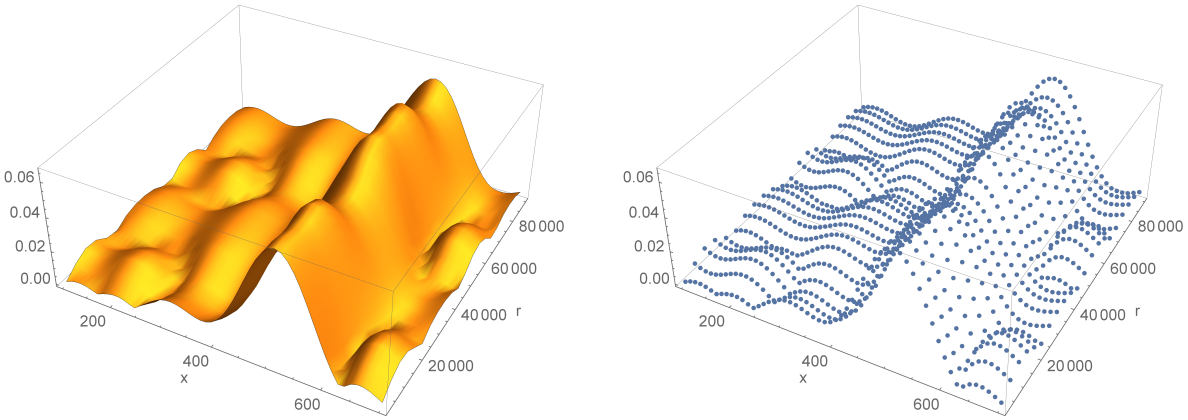

Figure 8. Reduction of the primeter data in order to avoid over-fitting. **Left**: reconstruction based on partial data, **Right**: resulting points.

8. Possible Applications

8.1. Process Simulation

The simulation is an important tool in both design and then subsequent maintenance of the system, especially if physical device is cumbersome, costly, or dangerous to use. The obvious method of simulating the engine temperature at specific rotational speed is to generate random numbers from a distribution specific for that temperature. The simplest method is the so-called acceptance-rejection method. In this method, a random number from distribution $f(x) \in [a, b]$ is generated as follows (in general using any dense random number generator, but typically uniform is used):

- generate random $x \in [a, b]$,
- generate $y \in [0, 1]$,
- if $y > f(x)$ then go to 1, otherwise the result is x.

We can easily obtain distribution for a specific parameter r. Using the example from Section 7.1, with r = 50,000 we obtain a distribution as in Figure 9.

If a random generation algorithm is used directly, the resulting process path would be highly irregular and perceived as not real. To avoid this, a simple filter can be used

$$\hat{T}_{i+1} = \tau \cdot \tilde{T} + (1 - \tau)\hat{T}_i, \tag{51}$$

where $\tilde{T}$ is the resulting new random number, $\hat{T}_{i+1}$ is the new process value, and $\hat{T}_i$ is the old process value. The result with parameter $\tau = 0.05$ can be seen in Figure 10.

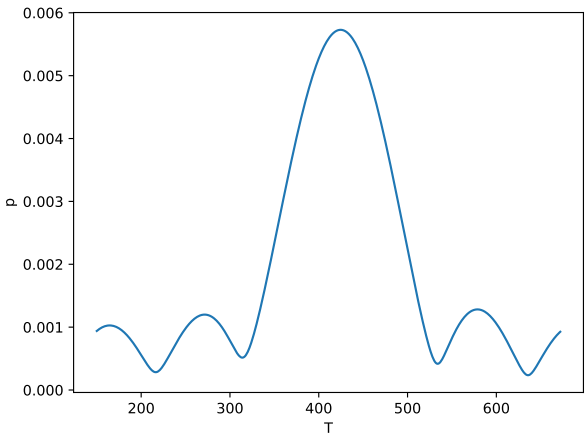

Figure 9. Probability for 50k rpm.

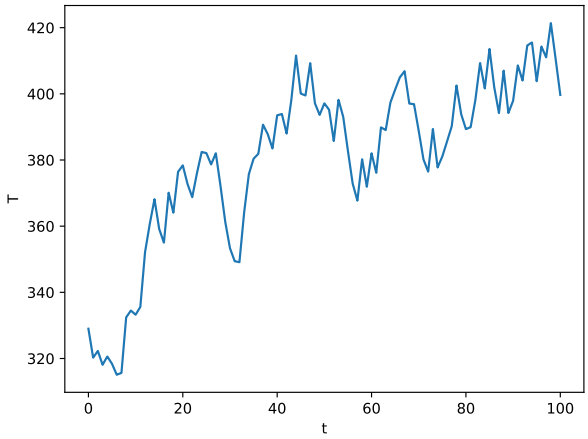

Figure 10. A result of smoothed simulation for 50k rpm.

8.2. Process Diagnostics

From the resulting function $f(r, T)$, for specific r, we can calculate expected the value of $\bar{T}$ and also some specified interval where $\alpha \in [0, 1]$ of realizations can be found—the equivalent of a confidence interval. If this parameter α is selected carefully, we can discern whether the actual pair r, T form the process in or outside it. Other similar diagnostic methods can be seen in [23,24].

9. Convergence of the Learning Procedure

In this section, we prove the convergence of the learning procedure in a simplified version similar to that described in Section 2.2, but without the discretization with respect to r. Then, we shall discuss additional conditions for the convergence of its discretized version.

Let $X(r), X_1(r), \ldots, X_n$ be independent, identically distributed random variables with parameter r, with p.d.f. $f(., r) \in L^2([R_1, R_2])$, where—for simplicity—we assume that n is the same for each r. Then, f has the representation

$$f(x, r) = \sum_{k=1}^{\infty} a_k(r)\, v_k(x), \tag{52}$$

which is convergent in the $L^2([S_1, S_2])$ norm, where

$$a_k(r) = \int_{S_1}^{S_2} v_k(x)\, f(x, r)\, dx = E_r(v_k(X(r))). \tag{53}$$

Then, we estimate $a_k(r)$'s as follows:

$$\hat{a}_k(r) = \frac{1}{n} \sum_{i=1}^{n} v_k(X_i(r)). \tag{54}$$

Lemma 4. *For every $r \in [R_1, R_2]$ $\hat{a_k}(r)$ is the unbiased estimator of $a_k(r)$.*

Proof. Indeed,

$$E_r(\hat{a_k}(r)) = \frac{1}{n} \sum_{i=1}^{n} E_r[v_k(X_i(r))] = a_k(r). \tag{55}$$

□

As an estimator of $f(., r)$, we take the truncated version of (52):

$$\tilde{f}_n(x, r) = \sum_{k=1}^{K(n)} \hat{a}_k(r)\, v_k(x), \tag{56}$$

where the truncation point K depends on n. It may also depend on r, but, for the asymptotic analysis, we take this simpler version.

The standard way of assessing the distance between f and its estimator $\tilde{f}_n$ is the mean integrated squared error ($MISE$). Notice, however, that, in our case, the MISE additionally depends on r, which is further denoted as $MISE(r)$. Thus, in order to have a global error, we consider the integrated $MISE(r)$, which is defined as follows:

$$IMISE_n = \int_{R_1}^{R_2} MISE_n(r)\, dr. \tag{57}$$

The $MISE(r)$ is defined as follows:

$$MISE_n(r) = E_r \int_{S_1}^{S_2} [f(x, r) - \tilde{f}_n(x, r)]^2 dx, \tag{58}$$

where the expectation w.r.t. $f(., r)$ concerns all $X_i(r)$'s that are present in (56).

Using the orthonormality of v_k's, we obtain:

$$MISE_n(r) = E_r \int_{S_1}^{S_2} [\sum_{k=1}^{K(n)} (a_k(r) - \hat{a}_k(r))^2\, v_k(x) - \sum_{k=K(n)+1}^{\infty} a_k(r)\, v_k(x)]^2 dx \tag{59}$$

Continuing (59), we obtain for each $r \in [R_1, R_2]$:

$$MISE_n(r) = \underbrace{E_r \left[\sum_{k=1}^{K(n)} (a_k(r) - \hat{a}_k(r))^2\right]}_{Var(r, K(n))} + \underbrace{\sum_{k=K(n)+1}^{\infty} a_k(r)^2}_{Bias^2(r, K(n))}. \tag{60}$$

It is known that that for squared integrable f we have: $\sum_{k=1}^{\infty} a_k^2(r) < \infty$. Thus, if $K(n) \to \infty$ when $n \to \infty$, then for every $r \in [R_1, R_2]$, we obtain:

$$Bias^2(r, K(n)) = \sum_{k=K(n)+1}^{\infty} a_k^2(r) \to 0 \text{ when } n \to \infty. \tag{61}$$

This result is not sufficient to prove the convergence of $IMISE_n$ to zero as $n \to \infty$. To this end, we need a stronger result, namely an upper bound on $Bias^2(r\, K(n))$, which is independent of r.

Lemma 5. *Let us assume that $\frac{\partial}{\partial x} f(x, r)$ exists and it is a continuous function of x in $[S_1, S_2]$ for each $r \in [R_1, R_2]$. Furthermore, there exists constant $0 < U < \infty$, which does not depend on r, and such that*

$$\forall\, x \in [S_1, S_2] \;\; \forall\, r \in [R_1, R_2] \quad \left|\frac{\partial}{\partial x} f(x, r)\right| \le U. \tag{62}$$

If $K(n) \to \infty$ when $n \to \infty$, then—for n sufficiently large, we have:

$$\forall\, r \in [R_1, R_2] \quad Bias^2(r, K(n)) \le \frac{U^2\, (S_2 - S_1)^2}{K(n)}. \tag{63}$$

Proof. It is known that

$$|a_k(r)| \le k^{-1} \int_{S_1}^{S_2} \left|\frac{\partial}{\partial x} f(x, r)\right| dx \le k^{-1}\, U\, (S_2 - S_1). \tag{64}$$

Thus, $Bias^2(r\, K(n)) \le U^2\, (S_2 - S_1)^2 \sum_{k=K(n)}^{\infty} k^{-2}$, which finishes the proof, since it is known that for sufficiently large $K(n)$ we have $\sum_{k=K(n)}^{\infty} k^{-2} = K^{-1}(n)$. □

For evaluating the variance component, we use Lemma 1:

$$Var(K(n),\, r) = \sum_{k=1}^{K(n)} E_r[(a_k(r) - \hat{a}_k(r))^2] = \sum_{k=1}^{K(n)} \underbrace{E_r(\hat{a}_k(r) - E_r\hat{a}_k(r))^2}_{Var_r(\hat{a}_k(r))}, \tag{65}$$

where $Var_r(.)$ is the variance of a r.v. having the p.d.f. $F(x,\, r)$.

Let us assume that the orthonormal and complete system v_k's is uniformly bounded, i.e., there exists p being a nonnegative integer and C_p such that

$$\sup_{x\in[S_1,\, S_2]} |v_k(x)| \,\leq\, C_p\, k^p, \quad k = 1,\, 2,\, \ldots\,. \tag{66}$$

Notice that (66) holds for the trigonometric system with $p = 0$

Lemma 6. *If* (66) *holds, then*

$$Var_r(K(n),\, r) \,\leq\, C_p\, \frac{K^{2\,p+1}(n)}{n} \tag{67}$$

Proof. $X_i(r)$'s are i.i.d. and (66) holds. Thus,

$$Var_r(\hat{a}_k(r)) = \frac{1}{n^2}\sum_{i=1}^{n} Var_r(v_k(X_i(r))) = \frac{1}{n} Var_r(v_k(X_1(r))) \leq \frac{C_p^2\, K^{2\,p}(n)}{n}. \tag{68}$$

Using this fact in (65) finishes the proof. □

Notice that the bound in (67) does not depend on r.

Theorem 2. *Let the assumptions of Lemmas 5 and 6 hold. If the sequence $K(n)$ is selected in such a way that the following conditions hold:*

$$K(n) \to \infty \text{ and } \frac{K^{2\,p+1}(n)}{n} \to 0 \text{ as } n \to \infty, \tag{69}$$

then the estimation error $IMISE_n \to 0$ as $n \to \infty$.

Proof. By Lemmas 2 and 6, we have uniform (in r) bounds for the variance and for the squared bias, respectively. Thus,

$$MISE_n(r) \,\leq\, C_p\, \frac{K^{2\,p+1}(n)}{n} + \frac{U^2\,(S_2 - S_1)^2}{K(n)}. \tag{70}$$

Hence,

$$IMISE_n \,\leq\, (R_2 - R_1)\, \left[C_p\, \frac{K^{2\,p+1}(n)}{n} + \frac{U^2\,(S_2 - S_1)^2}{K(n)}\right], \tag{71}$$

which finishes the proof by applying (69). □

Observe that for v_k's being the trigonometric system we have $p = 0$ and the r.h.s. of (71) is, roughly, of the form: $c_1\, K(n)/n + c_2/K(n)$, $c1 > 0$, $c_2 > 0$, which is minimized by $K(n) = c_3\, \sqrt{n}$ for a certain constant $c_3 > 0$. This implies that $IMISE_n = O(n^{-1/2})$. Notice that this rate is known (see [25]) to be the best possible when a bivariate, continuously differentiable regression function is estimated by any nonparametric method.

However, this convergence rate was obtained under an idealized assumption that we have observations of $X_j(r)$'s for each r. Further discussion of this topic is outside the scope of this paper. We only mention that in [17] it was proved that orthogonal series estimators of p.d.f.'s retain consistency under mild assumptions concerning grouped observations. In particular, in the notation of Section 3,

bin width Δ_x should depend on the number of observations d as follows: $\Delta_x(d) \to 0$ as $d \to \infty$, but in such a way that $\Delta_x^2(d)$ multiplied by the qube power of the number of spanning orthogonal terms should also approach zero. One can expect that, for the trigonometric series, Δ_r should also fulfill similar conditions.

10. Conclusions

In this paper, a method for the estimation of families of density functions is presented along with an efficient learning algorithm. It gives insight into the relation between probability density function and external factors that influence it. The method was used in a real-world problem to simulate and diagnose a jet turbine.

Additional, significant possible applications include the estimation of quality indexes for decision-making and optimal control, especially for repetitive and/or spatially distributed processes (see [26,27] for most recent contributions).

From the formal point of view, the method presented can easily be extended into a multidimensional case. Namely, if r is multivariate $\bar{r} = [r^{(1)}, r^{(2)}, \ldots, r^{(d)}]$, say, then it suffices to replace orthogonal system $T_j(r)$, $j = 1, 2, \ldots$ by the tensor product of T_j's, i.e., by all possible products of the following form:

$$T_{j_1}(r^{(1)})\, T_{j_2}(r^{(2)}) \cdot \ldots \cdot T_{j_d}(r^{(d)}), \tag{72}$$

where j_i takes all the values in $\{1, 2, \ldots\}$, $i = 1, 2, \ldots, d$. As a consequence, one has to replace all the sums over j by the multiple sums over j_i's. When T_{j_i}'s form the trigonometric system, then one can apply the multidimensional FFT algorithm. Clearly, a much larger number of observations is also necessary for a reliable estimation, but a family of estimated p.d.f.'s $\hat{f}(x, \bar{r})$ provides much more information than a regression function of $\bar{r}$.

Funding: This research received no external funding.

Acknowledgments: The author expresses his thanks to the anonymous reviewers for many helpful comments clarifying the presentation.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cencov, N.N. Evaluation of an unknown distribution density from observations. *Sov. Math. Dokl.* **1962**, *3*, 1559–1562.
2. Devroye, L. On the almost everywhere convergence of nonparametric regression function estimates. *Ann. Stat.* **1981**, *9*, 1310–1319. [CrossRef]
3. Györfi, L.; Kohler, M.; Krzyzak, A.; Walk, H. *A Distribution-Free Theory of Nonparametric Regression*; Springer Science & Business Media: New York, NY, USA; Berlin/Heidelberg, Germany, 2006
4. Parzen, E. Nonparametric statistical data modeling. *J. Am. Stat. Assoc.* **1979**, *74*, 105–121. [CrossRef]
5. Greblicki, W.; Danuta, R.; Rutkowski, L. An orthogonal series estimate of time-varying regression. *Ann. Inst. Stat. Math.* **1983**, *35*, 215–228. [CrossRef]
6. Rutkowski, L. Application of multiple Fourier series to identification of multivariable non-stationary systems. *Int. J. Syst. Sci.* **1989**, *20*, 1993–2002. [CrossRef]
7. Rutkowski, L. Real-time identification of time-varying systems by non-parametric algorithms based on Parzen kernels. *Int. J. Syst. Sci.* **1985**, *16*, 1123–1130. [CrossRef]
8. Duda, P.; Rutkowski, L.; Jaworski, M.; Rutkowska, D. On the Parzen kernel-based probability density function learning procedures over time-varying streaming data with applications to pattern classification. *IEEE Trans. Cybern.* **2020**, *50*, 1683–1696. [CrossRef]
9. Duda, P.; Jaworski, M.; Rutkowski, L. Convergent time-varying regression models for data streams: Tracking concept drift by the recursive Parzen-based generalized regression neural networks. *Int. J. Neural Syst.* **2018**, *28*, 1750048. [CrossRef] [PubMed]

10. Jaworski, M. Regression function and noise variance tracking methods for data streams with concept drift. *Int. J. Appl. Math. Comput. Sci.* **2018**, *28*, 559–567. [CrossRef]
11. Marron, J.S. A comparison of cross-validation techniques in density estimation. *Ann. Stat.* **1987**, 152–162. [CrossRef]
12. Bleuez, J.; Bosq, D. Conditions necessaires et suffisantes de convergence de l'estimateur de la densitepar la methode des fonctions orthogonales. *Rev. Roum. Math. Pures Appl.* **1979**, *24*, 869–886.
13. Devroye, L.; Györfi, L. *Nonparametric Density Estimation: The L1 View;* Wiley: New York, NY, USA, 1985.
14. Hall P. On the rate of convergence of orthogonal series density estimators. *J. R. Stat. Soc.* **1986**, *B48*, 115–122. [CrossRef]
15. Tarter, M.E.; Kronmal, R.A. An introduction to the implementation and theory of nonparametric density estimation. *Am. Stat.* **1976**, *30*, 105–112.
16. Hall, P. The influence of rounding errors on some nonparametric estimators of a density and its derivatives. *SIAM J. Appl. Math.* **1982**, *42*, 390–399. [CrossRef]
17. Rafajłowicz E. Consistency of Orthogoal Series Density Estimators Based on Grouped Observations. *IEEE Trans. Inf. Theory* **1997**, *10*, 283–285.
18. Rafajłowicz, E. Nonparametric orthogonal series estimators of regression: A class attaining the optimal convergence rate in L2. *Stat. Probab. Lett.* **1987**, *5*, 219–224. [CrossRef]
19. Rafajłowicz, E.; Skubalska-Rafajłowicz, E. FFT in calculation nonparameteric regression estimate based on trigonometric series. *Appl. Math. Comput. Sci.* **1993**, *3*, 713–720.
20. Hall, P.; Marron, J.S. Extent to which least-squares cross-validation minimises integrated square error in nonparametric density estimation. *Probab. Theory Relat. Fields* **1987**, *74*, 567–581. [CrossRef]
21. Kronmal, R.; Tarter, M. The estimation of probability densities and cumulatives by Fourier series methods. *J. Am. Stat. Assoc.* **1968**, *63*, 925–952.
22. Rao, K.; Kim, D.; Hwang, J. *Fast Fourier Transform-Algorithms and Applications*; Springer Science & Business Media: Dordrecht Heidelberg, London, The Netherlands; Heidelberg, Germany; London, UK; New York, NY, USA, 2011.
23. Gałkowski, T.; Krzyżak, A.; Filutowicz, Z. A New Approach to Detection of Changes in Multidimensional Patterns. *J. Artif. Intell. Soft Comput. Res.* **2020**, *10*, 125–136. [CrossRef]
24. Skubalska-Rafajłowicz, E. Random projections and Hotelling's T2 statistics for change detection in high-dimensional data streams. *Int. J. Appl. Math. Comput. Sci.* **2013**, *23*, 447–461. [CrossRef]
25. Stone, C.J. Optimal global rates of convergence for nonparametric regression. *Ann. Stat.* **1982**, 1040–1053. [CrossRef]
26. Mandra, S.; Galkowski, K.; Rauh, A.;Aschemann, H.; Rogers, E., Iterative Learning Control for a Class of Multivariable Distributed Systems With Experimental Validation. *IEEE Trans. Control Syst. Technol.* **2020**. [CrossRef]
27. Rafajłowicz, W.; Więckowski, J.; Moczko P.; Rafajłowicz, E. Iterative learning from suppressing vibrations in construction machinery using magnetorheological dampers. *Autom. Constr.* **2020**, *119*, 103326. [CrossRef]