WORKING PAPERS

N° 1770

September 2026

"A Note on Distribution and Density Estimation as a Regression Problem of Order Statistics"

Olivier Faugeras

Toulouse School of Economics

A NOTE ON DISTRIBUTION AND DENSITY ESTIMATION AS A REGRESSION PROBLEM OF ORDER STATISTICS.

OLIVIER P. FAUGERAS

Toulouse School of Economics, Université Toulouse Capitole, Institut de Mathématiques de Toulouse, France.

Abstract. We show how cumulative distribution function estimation (cdf) can be formulated as a regression problem, using order statistics. This results in a nonstandard nonparametric regression problem, with correlated but vanishing errors. Nonparametric regression estimators can then be leveraged to obtain corresponding estimators of the cdf and of its derivatives. We illustrate the approach, theoretically with Nadaraya-Watson regression estimator, and numerically with P-splines regression techniques. We discuss the issue of smoothing parameter selection for regression of correlated errors and show how to make use of known techniques from the regression framework to improve on the cdf and density estimation. Eventually, we draw some perspectives and show how the proposed framework could be extended to the multivariate setting.

1. Introduction and outline

Estimating the cumulative distribution function (cdf) and the corresponding probability density function are fundamental tasks in non-parametric statistics (Silverman, 1986, Scott, 2015, Wand and Jones, 1995). Traditionally, these are approached as unsupervised learning problems (Hastie, Tibshirani, and Friedman, 2009). In this note, we propose a novel perspective by demonstrating how univariate cdf estimation can be reformulated as a supervised non-linear regression problem. By framing distribution and density estimation within a supervised regression framework, our aim is to bridge these two paradigms, enabling the direct application of powerful regression methodologies to an unsupervised domain.

The core of the approach leverages the distributional properties of order statistics. By utilizing the quantile transform, evaluating the true, unknown cdf at the observed order statistics generates latent uniform order statistics. While these uniform variables are unobserved, their expectations are exactly known. By substituting these latent variables with their known expectations—a technique we term "forced observability"—we derive a regression framework where the sample order statistics act as the predictor variables.

E-mail address: olivier.faugeras@tse-fr.eu.

Date: September 4, 2026.

2020 *Mathematics Subject Classification.* 62G07, 62G08, 62G20, 62G30.

Key words and phrases. Cumulative distribution function estimation, density estimation, nonparametric regression, order statistics, P-splines, Nadaraya-Watson estimator, correlated errors.

This resulting model is highly non-standard compared to classical regression. Specifically, the response variable is deterministic, the design is random, and the errors are correlated with both themselves and the dependent variables. Crucially, however, these regression errors vanish asymptotically. This unique feature allows us to apply existing non-parametric regression estimators to successfully recover the cdf and, through its derivative, the density function.

The remainder of this paper is organized as follows: Section 2 formulates the baseline regression framework for univariate cdf estimation, highlights its fundamental differences from classical regression models, and reviews related literature. Section 3 provides theoretical validation via an asymptotic analysis of the classical Nadaraya-Watson (Nadaraya, 1964b, Watson, 1964) kernel estimator, establishing its pointwise and uniform consistency. Section 4 illustrates the approach practically and empirically using P-splines (Eilers and Marx, 1996, Eilers and Marx, 2010). This section addresses the critical challenge of smoothing parameter selection in the presence of correlated errors—advocating for the V-curve method—and demonstrates augmented techniques to improve density estimation, such as zero-slope correction and double penalization. We conclude in Section 5 and Section A offer supplementary discussions on related regression approaches to density estimation, propose an extension to the multivariate setting using distance-based order statistics, and outline open problems for future research, including robust regression and generalized least squares. All proofs are deferred to Appendix B.

2. Univariate cdf estimation as a regression problem

In this section, we formulate the estimation of a univariate cdf as a regression problem.

2.1. Basic regression framework.

Let $(X_1, \ldots, X_n)$ be an i.i.d. sample from a continuous univariate distribution F, with quantile function Q, and set $X_{(1)} \leq \ldots \leq X_{(n)}$ the corresponding order statistics[1]. In view of the quantile transform Theorem (see e.g. Shorack, 2000), one can always consider that the sample is generated from a set $(U_i)_{1\leq i\leq n}$ of i.i.d. latent variables, uniformly distributed on $[0, 1]$,

$$X_i = Q(U_i), \quad i = 1, \ldots, n.$$

Since F is continuous, the r.v. $U_{(1)} \leq \ldots \leq U_{(n)}$ defined as

$$U_{(i)} = F(X_{(i)}), \quad 1 \leq i \leq n \tag{1}$$

are the corresponding order statistics of $(U_1, \ldots, U_n)$. As the order statistics $(U_{(i)})$ are not observed, Equation (1) can not be directly used to formulate the problem of estimating F in a regression framework.

However, the distributional properties of the uniform order statistics are (well) known (see e.g. David and Nagaraja, 2003, Reiss, 1989). In particular, $\mathrm{E}[U_{(i)}] = i/(n+1)$. Therefore, and this is the basic idea of the paper, one can "force observability" of the unknown $(U_{(i)})$ by approximating them by their (known) expectation: define ϵ_i as the difference between the mean and the order statistics, viz.

$$\epsilon_i := i/(n+1) - U_{(i)}. \tag{2}$$

[1]We suppress the dependence of $X_{(i)}$ on n to simplify notation. Similarly for ϵ_i.

Then, (1) writes as

$$\frac{i}{n+1} = F(X_{(i)}) + \epsilon_i, \quad i = 1, \dots, n, \tag{3}$$

where the "noise" variables (ϵ_i) are centered, $\mathrm{E}[\epsilon_i] = 0$ with known covariance (see e.g. Durbin and Knott, 1972),

$$cov(\epsilon_i, \epsilon_j) = cov(U_{(i)}, U_{(j)}) = \frac{i(n+1-j)}{(n+2)(n+1)^2}, \quad \text{for } i \leq j. \tag{4}$$

Equation (3) now defines a (global) non-linear additive regression model (with correlated errors). It gives a natural formulation of the estimation of the cdf from a regression perspective. It suggests to use nonparametric regression estimators to obtain corresponding estimators of F and its derivatives. In particular, an estimate of the first derivative of F gives a regression-based estimator of the density of X.

Remark 1 (Intuition and link with sufficient statistics)**.** *Note that, for the reasoning leading to model (3), if one tries to apply the"forced observability" trick directly on the unordered statistics $U_i = F(X_i)$, this leads to a model*

$$y_i := EU_i = 1/2 = F(X_i) + \check{\varepsilon}_i,$$

where $\check{\varepsilon}_i := EU_i - U_i$ is the new (centred) "error". Thus, one gets the same response value $y_i = 1/2$ for all input observations X_i, which is not informative enough. In addition, each error $\check{\varepsilon}_i$ is free to vary independently across the range $[-1/2, 1/2]$ and thus is now an $O_p(1)$, i.e. too large (compared to (6) below).

In contrast, due to the order constraint induced by the sorting process, each $U_{(i)}$ can not exceed $U_{(i+1)}$, nor fall below $U_{(i-1)}$. Thus, $U_{(i)}$ is on average in an interval of expected length $2/(n+1)$, and each error ϵ_i of (2) is thus "locked" into a narrower interval than the interval $[-1/2, 1/2]$ of $\check{\varepsilon}_i$. This is also reflected in its variance: by (4),

$$Var(\epsilon_i) = \frac{i(n-i+1)}{(n+2)(n+1)^2} \leq \frac{1}{n+2},$$

which decreases with the sample size, compared to the fixed variance $1/12$ of $\check{\varepsilon}_i$.

This is to be expected, in view of the well-known fact that the order statistics are sufficient for estimating the cdf of a dominated distribution (see e.g. Ibragimov and Khasminskii, 1981 p. 13.). It thus appears necessary, both at an intuitive and at a theoretical level, to use the order statistics, before using the "forced observability" trick of Equation (1).

2.2. **Discussion.** We note some fundamental differences in (3) with standard regression models of the type

$$Y_i = r(X_i) + \epsilon_i. \tag{5}$$

1. the output variable $y_i = i/(n+1)$ is deterministic, and not random;
2. the dependent variables $(X_{(i)})$ are random (random design), and (positively) correlated;
3. the errors (ϵ_i) are correlated, both with themselves and with the dependent variables. In particular, one does not have $E[\epsilon_i | X_i] = 0$, as is required for the specification of a classical regression model (5), but $E[\epsilon_i | X_{(i)}] = \epsilon_i$. The latter is uniformly small in i, as $n \to \infty$ (see point 4. below).

Hence, it is not self-evident a-priori that the basic idea make sense. Fortunately, we have the following additional unusual feature on the regression error:

4. the error is not an $O_p(1)$ as is the case in standard regression models when $Var[\epsilon] = \sigma^2$ (homoscedastic case) or $Var[\epsilon] = \sigma^2(x)$ (heteroscedastic case), but is, uniformly in $1 \leq i \leq n$, an $O_p(n^{-1/2})$, as given by the following Lemma.

Lemma 2.1. *The set of errors r.v.* $\{\sqrt{n}\epsilon_i, 1 \leq i \leq n\}$ *is uniformly stochastically bounded in probability, viz.*

$$\max_{1\leq i\leq n} |\epsilon_i| = O_p\left(\frac{1}{\sqrt{n}}\right) \tag{6}$$

Remark 2. *By the same device, the law of the iterated logarithm for the empirical distribution function (Chung, 1949, Smirnoff, 1939,Finkelstein, 1971) entails*

$$\max_{1\leq i\leq n} |\epsilon_i| = O_{a.s}\left(\sqrt{\frac{\ln\ln}{n}}\right).$$

The fact that the error becomes asymptotically negligible will be key to show that the proposed regression estimator of the cdf is asymptotically consistent, in spite of the non-standard regression setting, as will be shown in Section 3.

Remark 3. *In the regression model (3), the full distribution of the errors* (ϵ_i) *is known. Indeed, the marginal distribution* $U_{(i)}$ *is* $\beta(i, n+1-i)$ *and the joint distribution of the order statistics* $(U_{(1)}, \ldots, U_{(n)})$ *is known to be the uniform distribution on the "ordered" (or first stick-breaking view, see Faugeras, 2026) representation of the probability simplex*

$$A_n := \{(x_1, \ldots, x_n) \in \mathbb{R}^n : 0 \leq x_1 \leq \ldots \leq x_n \leq 1\}.$$

That is to say the joint density of $(U_{(1)}, \ldots, U_{(n)})$ *is*

$$f_{(U_{(1)},\ldots,U_{(n)})}(x_1, \ldots, x_n) = n!\mathbb{1}_{A_n}(x_1, \ldots, x_n)$$

Hence, the density of $(\epsilon_1, \ldots, \epsilon_n)$ *writes*

$$f_{(\epsilon_1,\ldots,\epsilon_n)}(x_1, \ldots, x_n) = f_{(U_{(1)},\ldots,U_{(n)})}\left(x_1 + \frac{1}{n+1}, \ldots, x_n + \frac{n}{n+1}\right).$$

Having such complete knowledge of the distribution of the errors is in contrast with classical regression models, which do not inherently require a full distributional specification to estimate the regression coefficients. (The i.i.d. Gaussian error assumption is only required at the inference stage, for valid confidence intervals and hypothesis testing). However, we will not make use of the full knowledge of the distribution of the errors beyond second moment information.

Remark 4 (An alternative regression model)**.** *Lemma 2.1 suggests another approach for deriving a regression model to estimate the cdf* F*. It stems from the following straightforward observation: let* F_n *denotes the usual empirical cdf based on a sample of size* n*. Given that* $F_n(X_{(i)}) = i/n$*, the trivial decomposition* $F_n = F + F_n - F$ *applied at the* $i-$*th order statistics* $X_{(i)}$ *allows to write*

$$\frac{i}{n} = F(X_{(i)}) + \epsilon_i^*, \tag{7}$$

where the "error" is now the ϵ_i^* *of (17) in the proof of Lemma 2.1. Compared to model (3), the error* ϵ_i^* *is biased (but asymptotically unbiased) and also uniformly of order* $O_p(n^{-1/2})$*. Both models are asymptotically equivalent.*

2.3. **Related works.** The proposed regression framework (3 to estimate the cdf and its derivative seems to be new. See Servien, 2009 for an extensive bibliography on smooth estimation of the cdf. We review below the key methods to cdf smoothing proposed in the literature.

- One obvious approach to obtain a smooth nonparametric estimate the cdf is to integrate a density estimator, so that cdf estimation is closely related to density estimation. For the Parzen-Rosenblatt kernel density estimator, one obtains a cdf estimator by convolving the empirical cdf with a kernel k (Nadaraya, 1964a, Azzalini, 1981). This results in an estimator of the form
$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} L\left((x - X_i)/h\right),$$
where $L(x) = \int_{-\infty}^{x} k(t)dt$, is a distribution function. We refer to Yamato, 1972/73a, Yamato, 1972/73b, Winter, 1973, Fernholz, 1991, Chacón and Rodríguez-Casal, 2010, Chacón, Monfort, and Tenreiro, 2014 and the references therein for more details on kernel-smoothed estimates of the cdf.
- The method which appears to be the closest in spirit to the proposed approach is the one of Lejeune and Sarda, 1992, (see also Abdous and Bensaid, 2001), which minimises the weighted L_2 norm
$$J(a_0, \ldots, a_p; x) := \int_{-\infty}^{\infty} K\left(\frac{x-u}{h}\right)\left(F_n(u) - \sum_{k=0}^{p}\frac{a_k}{k!}(u-x)^k\right)^2 du$$
between the empirical cdf F_n and a polynomial approximation thereof of degree p, locally around x. It is thus based on a local polynomial approximation of the ecdf, and not on the regression representation (3).
- The proposed regression approach can be interpreted as a smoothing of the Q-Q plot (where the empirical quantile points of the uniform distribution are i/n, instead of the present $i/(n+1)$). Also, as a method based on order statistics, it has some similarities (but is different) with nearest-neighbours methods, which use the order statistics of the distances $||X_i - x||$ instead, see Biau and Devroye, 2015.

3. The Nadaraya-Watson estimate of the cdf

In this section, we validate theoretically—via an asymptotic analysis—the basic regression framework (3) of Section 2 for the estimation of the cdf, employing as non-parametric regression estimator the classical Nadaraya-Watson estimator. This gives the following estimator of F:

$$\hat{F}_0(x) := \frac{\sum_{i=1}^{n} i(n+1)^{-1} K_h(X_{(i)} - x)}{\sum_{i=1}^{n} K_h(X_{(i)} - x)}, \tag{8}$$

where K is the kernel function, h the bandwidth, and $K_h(x) := h^{-1}K(x/h)$

3.1. **A quick numerical simulation.** Before embarking on the asymptotic analysis, we run a quick simulation, using Mathematica, on a sample of size 200 from a $\mathcal{N}(0,1)$ distribution, in order to check if the basic regression framework (3) makes sense. The simulation is displayed in Figure 1. After playing a bit with the bandwidth, one gets an accurate estimator which corroborates numerically the basic

idea. More extensive simulations —notably using P-splines, will be developed in Section 4.

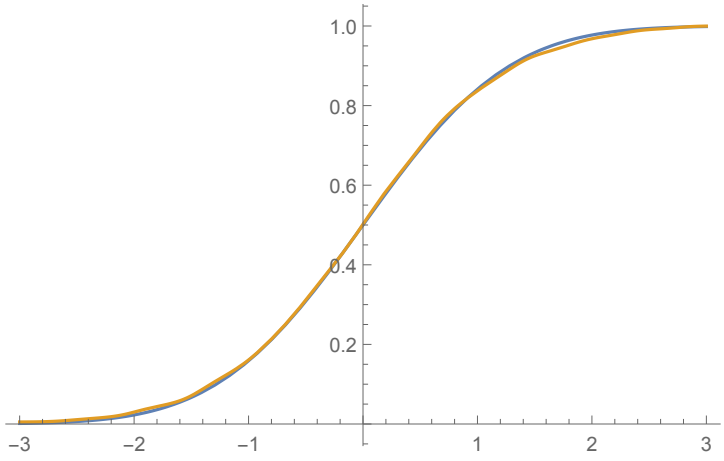

FIGURE 1. Nadaraya-Watson estimate of the cdf of a standard Gaussian. Blue: true cdf. Orange: estimate, $h = 0.2$.

3.2. **Preliminaries on kernel smoothing.** We begin by some preparations, setting up some notation and recalling some definitions. For a real number β, let $\lfloor \beta \rfloor$ denote the greatest integer strictly less than β. For $\ell \geq 1$ an integer, a function $K : \mathbb{R} \to \mathbb{R}$ is a kernel of order ℓ if the functions $u \mapsto u^j K(u)$, $j = 0, 1, \ldots, \ell$, are integrable and satisfy

$$\int K(u)du = 1, \quad \int u^j K(u)du = 0, \quad j = 1, \ldots, \ell.$$

The definition of the Hölder class of functions, see e.g. Tsybakov, 2004 p. 5, is given by:

Definition 3.1 (Hölder class). *Let I be an interval in $\mathbb{R}$, β and L be two positive numbers. The Hölder class $\mathcal{H}(\beta, L)$ on I is defined as the set of $\ell = \lfloor \beta \rfloor$ times differentiable functions $f : I \to \mathbb{R}$ whose derivative $f^{(\ell)}$ satisfies*

$$|f^{(\ell)}(x) - f^{(\ell)}(y)| \leq L|y - x|^{\beta - \ell}, \quad \forall x, y \in I.$$

One then has the following Taylor expansion with uniform control of the remainder.

Lemma 3.2 (Taylor for a function in a Hölder class)**.** *If $F \in \mathcal{H}(\beta, L)$ on $\mathbb{R}$, then, for $x \in I$, u s.t. $x + u \in I$,*

$$F(x + u) = \sum_{m=0}^{\ell} \frac{F^{(m)}(x)}{m!} u^m + \rho(x, u), \tag{9}$$

with $|\rho(x, u)| \leq L \frac{|u|^\beta}{\ell!}$.

The following proposition establishes the properties of the localized moments.

Proposition 3.3. *Let $q \in \mathbb{N}$, define $\mu_q(K) := \int x^q K(x)dx$ and*

$$s_q(x,h) := \frac{1}{nh}\sum_{i=1}^{n}(X_i - x)^q K\left(\frac{X_i - x}{h}\right).$$

Assume

i) *the kernel $K : \mathbb{R} \to \mathbb{R}$ is s.t.*

$$\mu_0(K) = 1, \quad K(-x) = K(x), \quad supp(K) = [-1, 1],$$

ii) *$h \downarrow 0$,*
iii) *$F \in \mathcal{H}(\beta, L)$, with $\beta > 1$.*

Then,

$$\mathrm{E}[s_q(x,h)] = \begin{cases} f(x)\mu_q(K)h^q + o(h^q) & \text{if } q \text{ is even,} \\ f'(x)\mu_{q+1}(K)h^{q+1} + o(h^{q+1}) & \text{if } q \text{ is odd,} \end{cases}$$

$$\mathrm{Var}[s_q(x,h)] \leq \frac{h^{2q}}{nh}\mu_{2q}(K^2)f(x)(1 + o(1)).$$

3.3. **Asymptotic analysis.** Several methods exist for investigating the asymptotic properties of the estimator (8). The conventional approach, which relies on direct computation of bias and variance, is rendered complex due to the correlation and non-identical distributions of the order statistics $(X_{(i)})$. Consequently, we adopt an alternative strategy in this study, leveraging an asymptotic decomposition to simplify the analysis. We have the following theorem establishing the pointwise consistency of the proposed estimator, with rates:

Theorem 3.4. *Assume*

i) *the kernel K is a symmetric probability density with support $[-1, 1]$.*
ii) *$h \downarrow 0$, $nh \to \infty$;*
iii) *$F \in \mathcal{H}(\beta, L)$, with $\beta > 1$ and $f(x) > 0$;*

Then,

$$\hat{F}_0(x) - F(x) = O(h^{\beta \wedge 2}) + O_p(n^{-1/2}),$$

where $a \wedge b = \min(a, b)$.

Theorem 3.4 reveals some interesting features:

i) First, it shows the saturation effect of the Nadaraya-Watson estimator: choosing as kernel a positive density function yields a kernel of order 1, at most. Hence, the approximation term in the decomposition can not take advantage of additional smoothness beyond the second derivative. This effect can also be explained by Korovkin's theorem (Korovkin, 1953, Altomare and Campiti, 1994): a kernel of order 1 for the Nadaraya-Watson estimator induces a positive linear operator of the observations, and the saturation class of positive linear operator contains functions of class $\mathcal{C}^2$.
ii) Contrary to the classical regression setting where the stochastic term $\hat{m}(x) - \mathrm{E}[\hat{m}(x)]$ of the Nadaraya-Watson estimator is of order $O(\frac{1}{\sqrt{nh}})$, there is here no influence of the bandwidth in the stochastic term $O_p(n^{-1/2})$. This is due to the negligible error (ϵ_i) of order $O_p(n^{-1/2})$, as given in Lemma 2.1.

As a consequence of Theorem 3.4, if the approximation term $O(h^{\beta \wedge 2})$ is negligible w.r.t. the stochastic term $O_p(n^{-1/2})$, one gets the same (parametric) rate of

convergence $n^{-1/2}$ as the empirical cdf, which is reassuring. This gives pointwise consistency in probability and in Mean Square Error (MSE) at the correct rate:

Corollary 3.5. *If, in addition, $h \sim n^{-\alpha}$, with $1/2 < \alpha < 1$, then*

i) $\hat{F}_0(x) = F(x) + O_p(n^{-1/2})$.
ii) $\mathrm{E}[(\hat{F}_0(x) - F(x))^2] = O(n^{-1})$.

Theorem 3.4 can be strengthened to uniform consistency, with a suitable boundedness assumption on the density and the derivative.

Corollary 3.6. *If the assumptions of Theorem 3.4 are satisfied and if, in addition, f, f' are bounded from above, then*

$$||\hat{F} - F||_\infty = O(h^{\beta \wedge 2}) + O_p(n^{-1/2}).$$

4. Nonparametric regression estimates of cdf and density functions: the case of P-Splines

Moving further than the Nadaraya-Watson estimator, the nonlinear regression formulation (3) of the cdf theoretically allows to estimate the cdf by any regression method. For example, as of April the first, 2025, the R-package caret (Kuhn and Max, 2008) lists 136 regression methods out of its 238 classification and regression techniques implemented in R (R Core Team, 2022). Among all such possible methods, we focus in this section on P-Splines, which are regarded by their authors as, quote, "the best smoother on the planet" Eilers and Marx, 2021. For our numerical illustrations, we rely on a simple simulated data model—specifically, a sample of $n = 200$ observations generated from a standard Gaussian distribution. This dataset acts as a recurring example, providing a coherent thread that runs through the experiments to highlight the method's key features. Our approach in this section is primarily empirical and practical, with the goal of conveying a clear impression of what the proposed regression framework (3) can and cannot achieve.

4.1. P-splines regression estimation of the cdf and density functions.

P-splines (Penalized B-splines), introduced by Eilers and Marx, 1996, are a flexible and computationally efficient non-parametric regression technique. Their design combines a rich B-spline basis on a relatively large number $K \approx 20$–50 of equidistant knots, regularised with a discrete roughness penalty, in order to avoid over-fitting and to circumvent the difficult problem of knot selection in spline smoothing.

More precisely, the P-spline estimator is obtained by minimizing the penalized least-squares criterion:

$$\widehat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \Big\{ \|\mathbf{y} - \mathbf{B}\boldsymbol{\alpha}\|^2 + \lambda\, \boldsymbol{\alpha}^\top \mathbf{D}_d^\top \mathbf{D}_d\, \boldsymbol{\alpha} \Big\},$$

where $\mathbf{y} \in \mathbb{R}^n$ is the vector of observed data outputs y_i, $\mathbf{B} \in \mathbb{R}^{n \times K}$ is the B-spline basis matrix of degree p evaluated at the data inputs x_i, $\boldsymbol{\alpha} \in \mathbb{R}^K$ the vector of spline coefficients, $\mathbf{D}_d$ is the matrix representation of the d-th order backward difference operator, and $\lambda \geq 0$ is a smoothing parameter. The fitted values are $\hat{\mathbf{y}} = \mathbf{B}\widehat{\boldsymbol{\alpha}}$, where the $\mathbf{B}$ spline basis matrix can possibly be recomputed on a finer grid. The method is applied to the regression model (3), with $x_i = X_{(i)}$ and $y_i = i/(n+1)$, $i = 1, \ldots, n$. Unless specified otherwise, we take $p = 3$ (cubic splines), $K = 50$ knots, and a grid of 1000 evaluation points.

4.2. **Selection of the smoothing parameter.** As for any smoothing method, the key issue is the selection of the smoothing parameter λ, which is typically done by cross-validation. However, it is known that standard (generalized) cross-validation techniques for bandwidth selection perform very badly when the errors are correlated, typically undersmoothing when the correlations are positive, as they misinterpret the correlated noise as signal, see e.g. Altman, 1990; Hart, 1991; Opsomer, Wang, and Yang, 2001 for kernel smoothing. For correlated errors, Frasso and Eilers, 2015 suggest to use a technique called V-curve: roughly speaking, it aims at finding the "knee", or point of maximum curvature/diminishing returns, of the L-curve— plotting the residuals $\log_{10}(\|\mathbf{y}-\mathbf{B}\boldsymbol{\alpha}\|^2)$ against the penalty $\log_{10}(\|\mathbf{D}_d\,\boldsymbol{\alpha}\|^2)$ in log-scales. The V-curve is obtained by plotting the norm of the the log-penalty and log-residual differences against the geometric means of two consecutive λ. The optimal λ is given by the minimal point of such V-curve, see Eilers and Marx, 1996 p. 55 for details and explanations. Note, that we had to implement such method manually as the V-curve smoothing method is not included in the JOPS R-package of Eilers and Marx, 1996.

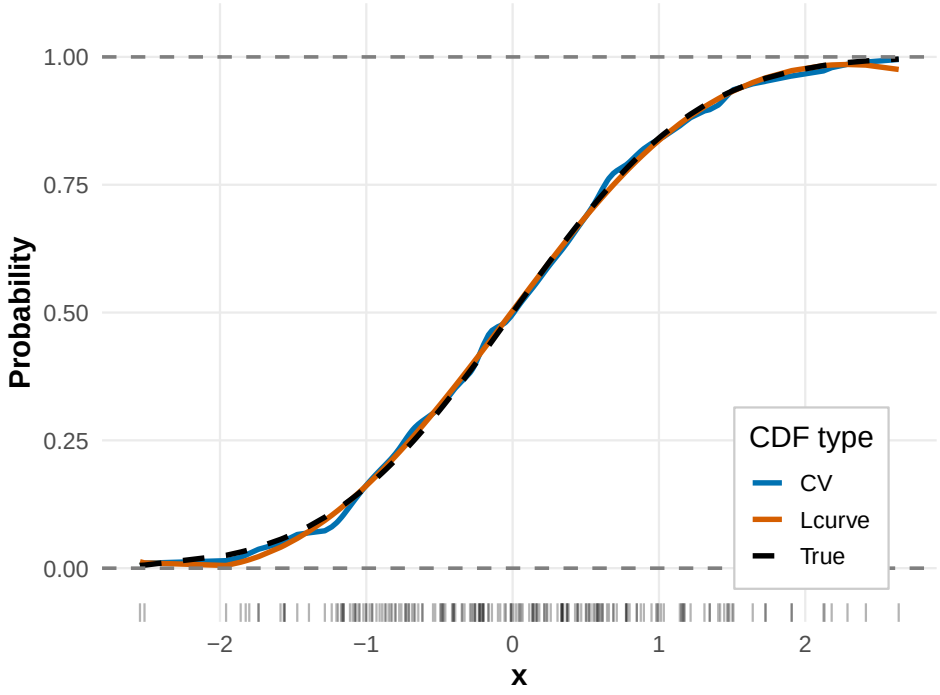

FIGURE 2. Comparison of the true cumulative distribution function (dashed black) and two P-splines regression estimates with smoothing parameter selected by cross-validation (blue curve, $\lambda = 0.015$) and the L/V-curve method (orange curve, $\lambda = 81$) for a sample of $n = 200$ $\mathcal{N}(0,1)$ Gaussian variables.

Figure 2 illustrates this issue of smoothing parameter selection for our basic example of 200 samples of a standard Gaussian distribution: the cubic P-spline regression estimator of the cdf with second-order difference penalization obtained by plain Cross-Validation (blue curve) selects a small $\lambda = 0.015$ and tends to overfit, closely matching the empirical ecdf (not shown). To the contrary, the corresponding P-spline estimator with V-curve method selects a larger parameter $\lambda = 81$ and produces a smoother curve (orange), which better matches the theoretical cdf (dashed black).

Since the true distribution is known in this simulation, one can calculate the Integrated L_2 distance $\int(\hat{F}(x) - F(x))^2 dx$ between the P-spline regression estimate $\hat{F}$ and the true Gaussian cdf F. This distance gives the exact generalization error of the estimate. Minimizing such distance gives the theoretically optimal smoothing parameter λ^*.

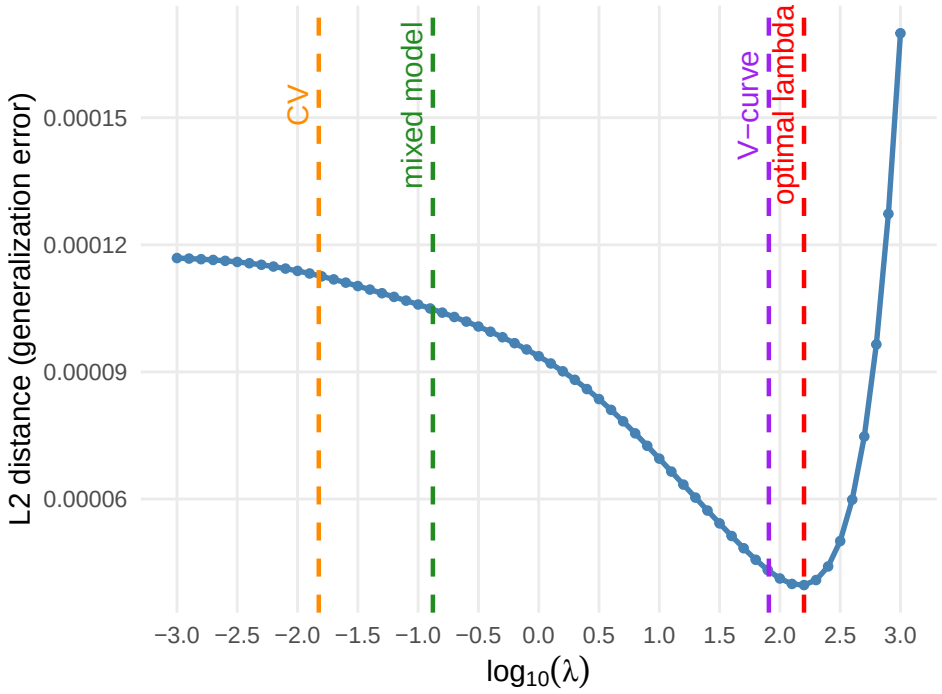

FIGURE 3. Comparison of integrated L_2 distance (generalization error) for P-splines w.r.t. to the smoothing parameter λ (in $\log_{10}$ scale). Theoretical optimal lambda (red) $\lambda^* = 158$, $\lambda = 0.015$ by Cross-Validation (orange), $\lambda = 81$ by the L/V-curve method (purple), $\lambda = 0.13$ by the Mixed Model method (green).

Figure 3 shows such generalization error (obtained by the Riemann sum on the evaluation grid) as a function of the smoothing parameter λ, for the simulated sample. The vertical lines show, respectively, the theoretically optimal lambda $\lambda^* = 158$ (green), the $\lambda = 0.015$ obtained by Cross-Validation (orange), and the $\lambda = 81$ by the L/V-curve method (purple): as mentioned before, plain CV selects

an overly small smoothing parameter, whereas the V-curve method, while still slightly undersmoothing, achieves the same order of magnitude as the theoretical optimum. Note that Eilers and Marx, 2010, bottom of p. 18, also propose a "hybrid" smoothing method which leverages the connection between P-splines with Mixed Models. We implemented their algorithm and obtained $\lambda = 0.13$ (green line in Figure 3), which improves on CV but remains inferior to their V-curve method.

4.3. **Density estimation.** P-splines allow for a simple computation of the derivative of the estimated function (see Eilers and Marx, 2021 p. 20), and thus provides for an estimate $\hat{f}$ of the density f of the data. Alternatively, the derivative can be computed by finite difference of the estimated cdf $\hat{F}$. We tried both methods, and found no visual difference, as long as the evaluation grid for the finite difference method is fine enough (≥ 1000 points).

Figure 4 shows the resulting estimate of the density. The P-spline regression estimate of the density with smoothing parameter picked by the previous V-curve selection technique (orange, top panel) is slightly wiggly, being dragged upwards where data points cluster, as shown on the corresponding histogram. For comparison, we computed a classical kernel density estimate (dotted blue curve) with Epanechnikov kernel and plug-in bandwidth selection (Method "SJ" in R-command "density"). The bottom panel shows the corresponding squared error $(f(x_i) - \hat{f}(x_i))^2$ at the corresponding locations x_i.

These visual results are to some degree a bit disappointing and are confirmed by the quantitive results of Table 1, where we computed the L_2 Integrated Square Error (ISE), $\sum_{x_j}(f(x_j) - \hat{f}(x_j))^2$, and the L_1 Integrated Absolute error (IAE), $\sum_{x_j}|f(x_j) - \hat{f}(x_j)|$, where x_j runs on the evaluation grid of 1000 points: the first two lines of Table 1 show that the regression P-spline estimate of the density is dominated by the kernel estimator.

TABLE 1. Integrated Square Error (ISE) and Integrated Absolute Error (IAE) for Density Estimation.

Method	ISE	IAE
Kernel smoothing	0.001 396	0.067 67
P-spline regression derivative	0.001 876	0.076 544
P-spline with zero slope correction	0.001 292	0.067 930
P-spline with double penalty	**0.001196**	**0.063526**

4.4. **Augmented regression P-spline density estimate with slope correction.** One can leverage the flexibility of the P-spline regression smoothing approach to improve on the somewhat disappointing results of the preceding section regarding density estimation. In particular, one can adapt to the special characteristics of the curve being estimated— that is a cdf is a non-decreasing function ranging from 0 to 1 and a density must be non-negative, by incorporating these shape constraints to the penalty. Indeed, there is a priori no guarantee that a P-splines density estimate (computed by finite difference or by the spline formula) will be a bona fide density— that is that the estimated derivative $\hat{f}$ remain positive, and

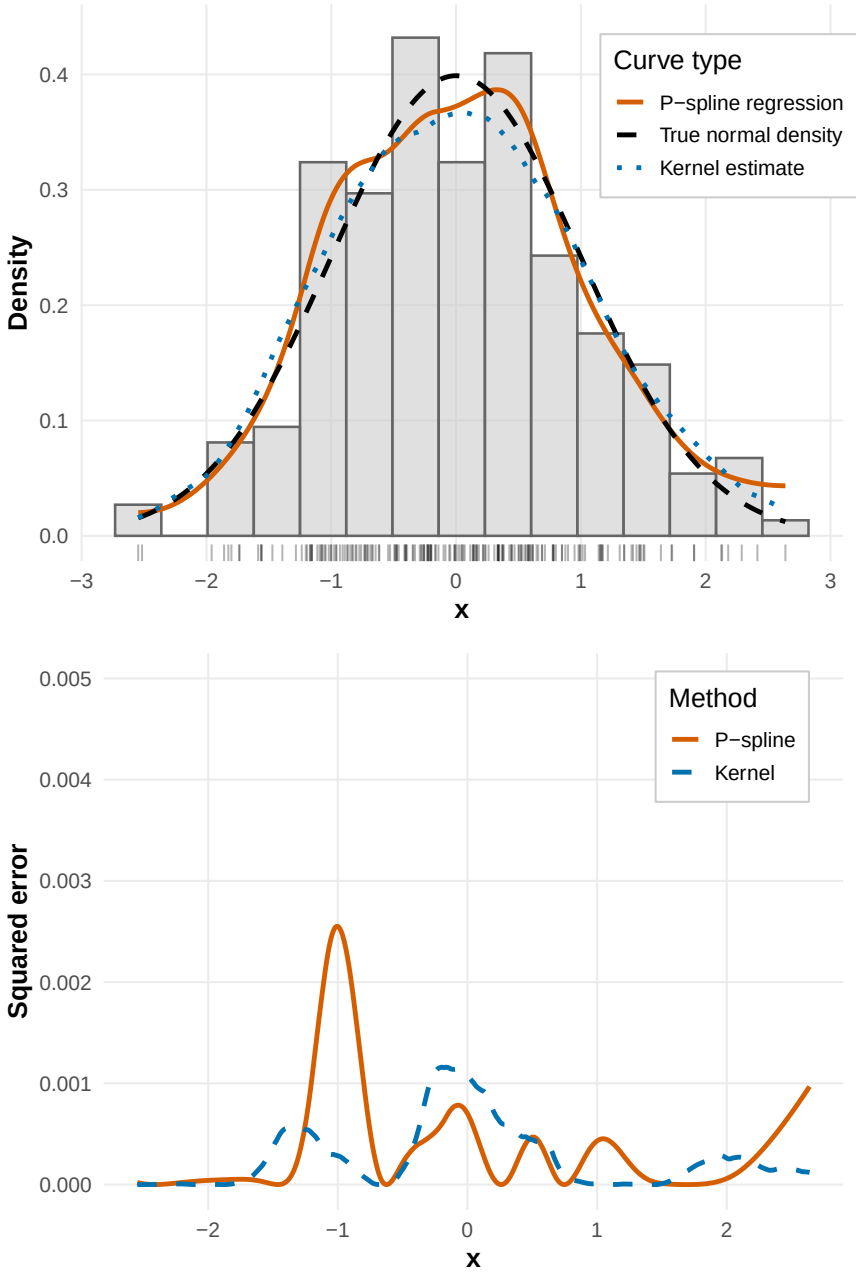

FIGURE 4. Comparison of the true density function (dashed black), the P-spline regression estimate with L/V-curve selection (orange), and a classical kernel estimate (dotted/dashed blue). Density estimates (top) and their squared errors (bottom).

we noticed that increasing the degree and/or difference order increases flexibility which typically induces a density estimate crossing the 0 level at both ends.

A suggestion, given by Eilers and Marx, 2021 Section 8.7 p. 148, to enforce flat asymptotes of the cdf at both endpoints, is to penalize its derivative at both ends. This requires selection of two left/right cut-off values $x_l < x_r$, which indicates where one must penalize the derivatives (for $x < x_l$ and for $x > x_r$). We can then use the above-mentioned defect of unconstrained P-spline derivative estimate— that of having a density estimate with negative values, to our advantage: the idea is to use such a defective P-spline derivative as a pilot, indicating where it crosses the 0 line, in order to select the left and right cut-off values to be used later in a final estimate.

Overall, this result in a regression P-spline density estimate with zero slope correction procedure conducted in two steps:

(1) first, a pilot P-spline estimate of the cdf with increased degree 4 and difference order 3 is estimated (with λ selected by the previous V-curve method). This yields an estimate escaping the $[0, 1]$ range. The left and right cut-off values $x_l < x_r$ are selected as the first x value where the pilot derivative estimate is > 0, respectively the last value where the pilot derivative estimate is > 0.
(2) Second, a P-spline estimate of the cdf with regular degree 3 and difference order 2 is estimated, with two extra penalties added, $\kappa_l \|\tilde{B}_l D_1 \boldsymbol{\alpha}\|^2$, and $\kappa_r \|\tilde{B}_r D_1 \boldsymbol{\alpha}\|^2$, where κ_l, κ_r are two large numbers (1000 in our simulations), $\tilde{B}_l$ and $\tilde{B}_r$ are the splines derivative basis matrices constructed on $x < x_l$ and $x > x_r$. These added penalties enforce the constraint that the derivatives of the cdf should be zero at both ends, before x_l and after x_r. The zero-slope corrected density estimate is then obtained by finite difference of this second P-spline estimate of the cdf.

Figure 5 shows the resulting zero-slope corrected density estimator: the P-spline pilot density estimate (dashed orange, top) crosses the zero level at $x_l \approx -2.20$ and $x_r \approx 2.28$. These values are then fed into the zero-slope corrected density estimator (blue) to penalize its derivative on both ends. This result in a density estimate which improves on both the plain P-spline density estimate of Figure 4, and on the kernel estimate (dotted blue), especially in the middle part of the density. The lower panel shows the squared error, which is now flatter, especially in the 95% percentile range. Table 1, line 3, confirms this impression, with improved ISE w.r.t. to kernel and plain P-spline smoothers, and IAE matching the kernel's one within two digits.

4.5. **Augmented regression for density estimation 2: the double penalty method.** Eilers and Marx, 2010 Figure 14, Section 13, are interested in estimating impulse response functions, which are characterized by a flat baseline with a sharp peak. They report improved estimation with a double penalty

$$\lambda \|\mathbf{D}_2 \boldsymbol{\alpha}\|^2 + 2\sqrt{\lambda} \|\mathbf{D}_1 \boldsymbol{\alpha}\|^2. \tag{10}$$

They explain that such combination of second and first order can prevent negative excursions and are useful when the curve has a flat baseline, that is zero asymptotes.

We tested such combination of penalties to our regression framework (3). The resulting density estimate is displayed in Figure 6 and Table 1 (last line). As there is no analogue of the V-curve method for such double penalty method, we selected

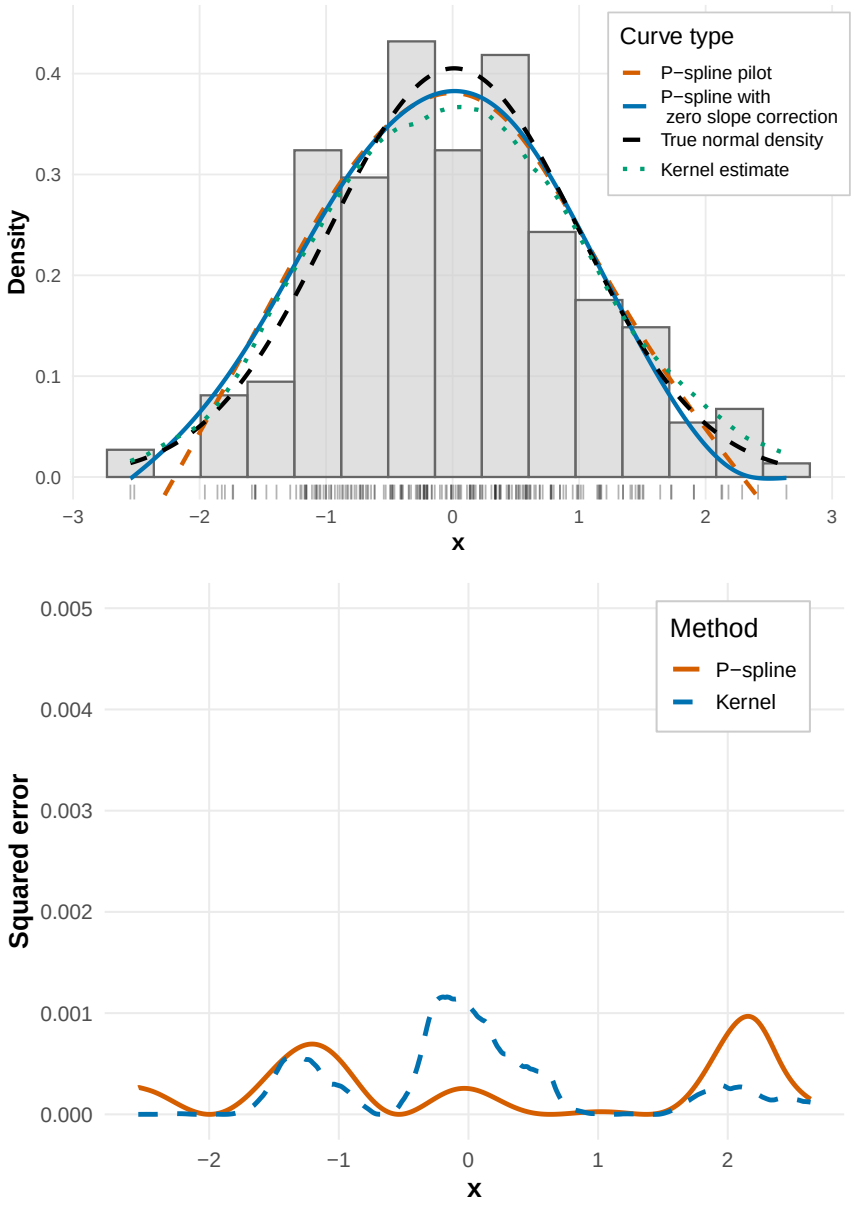

FIGURE 5. Density estimation for the P-spline regression estimate (with L-curve selection) and zero slope correction.

the smoothing parameter using the previous V-Curve method (thus for a P-splines with sole second order penalty), and plugged the obtained $\lambda = 81$ value in the double penalty (10). The idea is that by adding the $2\sqrt{\lambda}\|\mathbf{D}_1\boldsymbol{\alpha}\|^2$ term, one aims at correcting for the slight undersmoothing of the V-curve method that was reported in Figure 3.

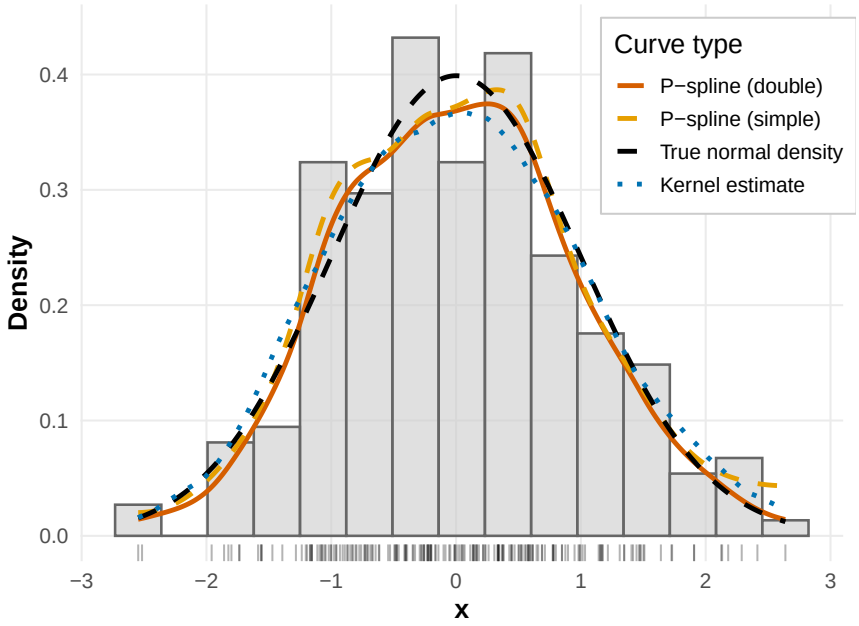

FIGURE 6. Comparison of the true density function (dashed black), the P-spline regression estimate with double penalty (orange-red), the P-spline regression estimate with simple penalty (dashed orange) and a classical kernel estimate (dotted blue).

The double penalty P-spline density estimator (solid dark orange curve) visually improves upon its single penalty counterpart (dashed orange) in Figure 6, and gets the overall best ISE and IAE in Table 1.

5. Conclusion

Leveraging the known distribution of the uniform order statistics, we have expressed the cdf as the regression/response function of the sample order statistics, with correlated but vanishing errors. This yields a non-standard regression model. Overall, we report mixed results: on the theoretical side, the framework is appealing as it allows to automatically transform any nonparametric regression estimation method into a corresponding cdf and (through its derivative) density estimation counterpart. It enables to view such unsupervised problem into a supervised one, potentially unlocking the vast literature on regression estimation to be applied on distribution and density estimation. Asymptotic results of Section 3, showing consistency at the correct rate of the Nadaraya-Watson based regression estimator of the cdf, validate the framework theoretically.

On the other hand, the non-standard nature of the regression problem, in particular the fact that the errors are correlated, make the choice of an appropriate smoothing parameter more difficult in practice, especially for the derivative estimation, that is the density. We illustrated in Section 4 how the flexibility of regression P-splines can be reused in the present framework to improve on the bandwidth selection process (the L/V-curve method) and incorporate a priori information on

the cdf via penalization. In particular, we showed how known techniques from the spline regression literature (slope correction by penalization and double penalization) can be tailored to the proposed framework and proposed two augmented density estimators.

For space constraints, we have limited ourselves to the univariate case, and to the Nadaraya-Watson and the P-splines methods, with the objective of showing how to bridge the gap, theoretically and empirically, between supervised and unsupervised learning methods. In order to stimulate further research and discussion, we close this note by adding supplementary discussions on i) related regression approaches in density estimation, ii) extension to the multivariate case and iii) further improvements attempts and open problems based on regression techniques.

Appendix A. Supplementary discussions and perspectives

A.1. **Related regression approaches in density estimation.** Related works formulating unsupervised problems as supervised ones are to be found in density estimation. In particular, the following approaches formulate density estimation into a regression framework.

(1) The binning method (apparently originating from Lindsey, 1974b, Lindsey, 1974a) turns the density estimation problem into a regression framework. Roughly speaking, since a histogram is a series of counts, a Poisson regression model can be applied to smooth it, owing to the law of rare events. More precisely, one partition an interval containing all n observations into N subintervals $\{I_k, k = 1, \dots, N\}$ of equal length δ. For each of the intervals I_k, set x_k the center of I_k and y_k the proportion of the data falling in the interval I_k divided by δ. Then, $n\delta y_k$ is binomially distributed with parameters n and $p_k := \int_{[x_k-\delta/2, x_k+\delta/2]} f_X(x)dx \approx \delta p_k$. Hence, estimation of the density f can be viewed as a nonparametric heteroscedastic regression problem, with corresponding data $\{(x_k, y_k), k = 1, \dots, N\}$.

As the Binomial $n\delta y_k$ can heuristically be approximated by a Poisson $\mathcal{P}(np_k)$, it make senses to use an Anscombe transform (Anscombe, 1948) $y \mapsto 2\sqrt{y + 3/8}$, which transforms a random variable with a Poisson distribution into one with an approximately standard Gaussian distribution. This suggests to use as data $\{(x_k, y_k^*), k = 1, \dots, N\}$ with

$$y_k^* = 2\sqrt{n\delta y_k + 3/8}$$

to estimate $2\sqrt{f(x) + 3/8}$, from which one can recover $f(x)$. See e.g. Fan and Gijbels, 1996 for local polynomial density estimation using such binning method.

(2) Local likelihood methods (Loader, 1996,Hjort and Jones, 1996, Loader, 1999) use local polynomial regression methods to estimate the density. They are based on minimizing a localized version of the log-likelihood

$$L(f; x) := \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \log f(X_i) - n \int K\left(\frac{x - u}{h}\right) f(u)du,$$

with a local polynomial approximation for $\log f(u) \approx \sum_{k=0}^{p} \frac{a_k}{k!}(u - x)^k$, in a neighbourhood of x. The local likelihood density estimate is then defined as $\hat{f}(x) = \exp \hat{a}_0$, where $\hat{a}_0$ is the value of a_0 which minimizes the above approximate local likelihood.

(3) Bos and Schmidt-Hieber, 2024 formulates density estimation as a regression problem using a data-splitting technique: in a first step, an under-smoothed (i.e. with a small bandwidth) kernel density estimate $\hat{f}_{\mathrm{KDE}}$ is computed from a first half of the sample, i.e. on $(\mathbf{X}_i)$, $i = 1, \ldots, n$, for a sample of size $2n$. In a second step, synthetic response variables

$$Y_i = \hat{f}_{\mathrm{KDE}}(\mathbf{X}_i), \quad n+1 \leq i \leq 2n$$

are constructed by applying the kernel density estimate of the first step to the other half of the sample. These response variables Y_i interprets as noisy versions of $f(\mathbf{X}_i)$, where f is the unknown density of $\mathbf{X}$. This corresponds to the regression model

$$Y_i = f(\mathbf{X}_i) + \epsilon_i, \quad n+1 \leq i \leq 2n$$

for the augmented data $(\mathbf{X}_i, Y_i)$ with the unknown density as regression function, and where $\epsilon_i := Y_i - f(\mathbf{X}_i)$ are the errors. These errors are correlated, since all data points $(\mathbf{X}_i, Y_i)$ are dependent on the preliminary kernel density estimator $\hat{f}_{KDE}$. A non-parametric regression estimate based on least squares minimization over a function class $\mathcal{F}$ is then used in the second estimate to produce the final estimator $\hat{f}$, viz.

$$\hat{f}_n \in \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - f(\mathbf{X}_i)\right)^2$$

Bos and Schmidt-Hieber, 2024 choose deep ReLu networks as functional class, in view of its ability to avoid the curse of dimensionality when the regression function has a composition structure and use Poissonization techniques to deal with the dependence at two different data points.

(4) On the theoretical side, it is known, since the seminal results of Nussbaum, 1996 (see also Genon-Catalot, Laredo, and Nussbaum, 2002, Ray and Schmidt-Hieber, 2016, Ray and Schmidt-Hieber, 2019), that density estimation on the unit interval is asymptotically equivalent to a Gaussian white noise experiment, provided the densities have Hölder smoothness larger than $1/2$ and are uniformly bounded away from zero.

A.2. **Multivariate extension.** The univariate framework of Section 2 exploits the total order of $(\mathbb{R}, \leq)$ to define sample order statistics. In the multivariate case $\mathbf{X} \in \mathbb{R}^d$, $d > 1$, the absence of a canonical total order hinders a direct generalization. Several remedies exist in the literature, including depth functions (Tukey, 1975; Zuo and Serfling, 2000; Mosler, 2013; Weller and Eddy, 2015), multivariate concomitants (Arnold, Castillo, and Sarabia, 2009a, 2009b), and optimal transport (Chernozhukov et al., 2017).

We indicate how a simple device of Biau and Devroye, 2015 can be used to adapt multivariate density estimation to the proposed regression framework (3) of Section 2: for a fixed $\mathbf{x} \in \mathbb{R}^d$, set

$$Y := Y(\mathbf{x}) := \|\mathbf{X} - \mathbf{x}\|^d,$$

reducing the multivariate $\mathbf{X}$ to the univariate Y. The cdf F_Y of Y relates to the density $f_{\mathbf{X}}$ of $\mathbf{X}$ via

$$F_Y(y) := \mathbb{P}(Y \leq y) = \mathbb{P}(\mathbf{X} \in B(\mathbf{x}, y^{1/d})) = \int_{B(\mathbf{x}, y^{1/d})} f_{\mathbf{X}}(\mathbf{t})\, d\mathbf{t}, \tag{11}$$

where $B(\mathbf{x}, r)$ denotes the closed Euclidean ball of radius r. By Lebesgue's differentiation theorem (Theorem 20.18 in Biau and Devroye, 2015), at every Lebesgue point $\mathbf{x}$ of $f_{\mathbf{X}}$,

$$\frac{\int_{B(\mathbf{x},r)} f_{\mathbf{X}}(\mathbf{t})\, d\mathbf{t}}{\lambda^d(B(\mathbf{x}, r))} \to f_{\mathbf{X}}(\mathbf{x}), \quad r \downarrow 0. \tag{12}$$

Translation invariance and scaling of the Lebesgue measure give $\lambda^d(B(\mathbf{x}, r)) = r^d V_d$, where V_d is the volume of the unit Euclidean ball. Combining (12) with (11) yields

$$f_Y(0) = \lim_{y \downarrow 0} \frac{F_Y(y)}{y} = V_d f_{\mathbf{X}}(\mathbf{x}). \tag{13}$$

Consequently, $f_{\mathbf{X}}(\mathbf{x})$ can be estimated from the derivative of F_Y at 0, using the order-statistic regression of Section 2. Given a sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$, let $Y_{(1)} \leq \cdots \leq Y_{(n)}$ be the order statistics of $Y_i = \|\mathbf{X}_i - \mathbf{x}\|^d$. The regression model corresponding to (3) is

$$\frac{i}{n+1} = F_Y(Y_{(i)}) + \epsilon_i, \quad i = 1, \ldots, n, \tag{14}$$

with ϵ_i as defined in Section 2. This estimation can be parallelized across different $\mathbf{x}$.

A.3. **Other avenues for possible improvements.** Here, we present further refinement attempts and discuss open challenges to advance the proposed framework, outlining potential directions for future research.

A.3.1. *Regression with correlated errors.* Hitherto, we have ignored the covariance structure of the errors. It is known in the theory of Generalised Least Squares (see e.g. Aitken, 1936, Kariya and Kurata, 2004), that the Best Linear Unbiased Estimator (BLUE) in a linear regression model with correlated errors is obtained by *weighted* least squares, with weight matrix W the precision matrix $\mathbf{\Sigma}^{-1}$, that is the inverse of the covariance matrix of the error. Such general statistical consideration suggest to use this information in our setting, even more that the covariance matrix $\mathbf{\Sigma} := \mathrm{cov}(\boldsymbol{\epsilon})$ of the errors (ϵ_i) is known exactly and given by (4). In addition, according to Durbin and Knott, 1972, $\mathbf{\Sigma}^{-1}$ can be computed in closed form and has the following triangular Toeplitz form:

$$\mathbf{\Sigma}^{-1} = (n+1)(n+2) \begin{bmatrix} 2 & -1 & 0 & 0 & \ldots & 0 \\ -1 & 2 & -1 & 0 & \ldots & 0 \\ 0 & -1 & 2 & -1 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ldots & 0 & -1 & 2 & -1 \\ 0 & \ldots & \ldots & 0 & -1 & 2 \end{bmatrix}. \tag{15}$$

However, numerical experiments (not shown) proved disappointing, with worse performance for the estimation of the cdf and increasing wiggliness for the estimated density with weighted P-splines or weighted local polynomial estimation with (15) than without weights. We nonetheless report a slight improvement using a diagonal matrix of weights given by the inverse of the variances, viz. $W := \mathrm{diag}(1/\mathrm{Var}(\epsilon_i))$.

A.3.2. *Robust regression.* In the regression model (3), the error ϵ_i is distributed as $\beta(i, n+1-i)$, and is thus quantitatively different from a Gaussian distribution, especially for i close to 1 and n. One may thus enquire whether it would make sense to try to improve on the estimator by using robust regression methods (Hampel et al., 1986, Maronna, Martin, and Yohai, 2006, Rousseeuw and Leroy, 1987, Bloomfield and Steiger, 1983, Riazoshams, Midi, and Ghilagaber, 2018) which are less sensitive to classical Gaussian regression assumptions.

Theoretically, a rationale could be described as follows: considering Least absolute deviation (L_1 regression), the median of a Beta distribution does not have a closed form in general, involving the regularized incomplete beta function. However, Payton, Young, and Young, 1989 Theorem 1 gives bounds between the mode, mean and median of a Beta distribution. This gives for the median m_i of the error ϵ_i, the following, for $n > 2$,

$$0 < \frac{i}{n+1} - m_i < \frac{n+1-2i}{(n-1)(n+1)}, \quad 1 < i \leq \lfloor n/2 \rfloor$$

and similarly, but with the order reversed, for $\lfloor n/2 \rfloor < i < n$. Hence, this shows that the median is close to the mean and that using median regression (which estimates the median) could be a reasonable surrogate for estimating the mean. This open problem is left for further research.

Appendix B. Proofs

Proof of Lemma 2.1. Decompose

$$\epsilon_i = \frac{i}{n+1} - \frac{i}{n} + \frac{i}{n} - U_{(i)} = \frac{-i}{n(n+1)} + \epsilon_i^*, \tag{16}$$

where we have set

$$\epsilon_i^* := \frac{i}{n} - U_{(i)} = F_n(X_{(i)}) - F(X_{(i)}), \tag{17}$$

and $F_n(x) = n^{-1} \sum_{i=1}^n \mathbb{1}_{X_i \leq x}$ denotes the usual empirical cdf. Since the absolute value of the first term in (16) is bounded by $(n+1)^{-1}$, the triangle inequality entails

$$\max_{1 \leq i \leq n} |\epsilon_i| \leq \frac{1}{n+1} + \|F_n - F\|_\infty.$$

Hence, by the Dvoretzky–Kiefer–Wolfowitz–Massart inequality (Dvoretzky, Kiefer, and Wolfowitz, 1956, Massart, 1990), one has that

$$\max_{1 \leq i \leq n} |\epsilon_i| = O_p\left(\frac{1}{\sqrt{n}}\right).$$

□

Proof of Lemma 3.2. By Taylor Formula with Lagrange remainder, there exists $0 \leq \tau \leq 1$ s.t.

$$\begin{aligned} F(x+u) &= \sum_{m=0}^{\ell-1} \frac{F^{(m)}(x)}{m!} u^m + \frac{u^\ell}{\ell!} f^{(\ell)}(x + \tau u) \\ &= \sum_{m=0}^{\ell} \frac{F^{(m)}(x)}{m!} u^m + \frac{u^\ell}{\ell!} (f^{(\ell)}(x + \tau u) - f^{(\ell)}(x)) \end{aligned}$$

Setting $\rho(x,u) = \frac{u^\ell}{\ell!}(f^{(\ell)}(x+\tau u) - f^{(\ell)}(x))$ and using the assumption yields

$$|\rho(x,u)| \leq \frac{|u|^\ell}{\ell!} L|u|^{\beta-\ell} = L\frac{|u|^\beta}{\ell!}.$$

□

Proof of Proposition 3.3.

$$\mathrm{E}[s_q(x,h)] = \frac{1}{h}\,\mathrm{E}\left[(X-x)^q K\left(\frac{X-x}{h}\right)\right] = h^q \int u^q K(u) f(x+uh)du$$

If q is even, $\mu_q(K) \neq 0$. $\beta > 1$ entails continuity of f at x. Thus,

$$\mathrm{E}[s_q(x,h)] = f(x)\mu_q(K)h^q(1+o(1)).$$

If q is odd, symmetry of K entails $\mu_q(K) = 0$. A Taylor expansion yields

$$\begin{aligned}\mathrm{E}[s_q(x,h)] &= h^q \int u^q K(u)[f(x+uh) - f(x)]du \\ &= h^{q+1}\mu_{q+1)}(K)f'(x) + o(h^{q+1}).\end{aligned}$$

$$\begin{aligned}\mathrm{Var}[s_q(x,h)] &= \frac{n}{n^2h^2}\,\mathrm{Var}\left[(W-x)^q K\left(\frac{X-x}{h}\right)\right] \\ &\leq \frac{1}{nh^2}\,\mathrm{E}\left[(W-x)^{2q} K^2\left(\frac{X-x}{h}\right)\right] \\ &= \frac{h^{2q}}{nh}\int u^{2q}K^2(u)f(x+hu)du\end{aligned}$$

Therefore, by continuity of f,

$$\mathrm{Var}[s_q(x,h)] \leq \frac{h^{2q}}{nh}\mu_{2q}(K^2)f(x)(1+o(1)).$$

□

Proof of Theorem 3.4. Plugging (3) and (9) in the definition (8) of the estimator yields the decomposition

$$\begin{aligned}\hat{F}_0(x) = F(x) &+ \sum_{m=1}^{\ell} \frac{F^{(m)}(x)}{m!}\frac{\sum_{i=1}^n (X_i-x)^m K((X_i-x)/h))}{\sum_{i=1}^n K((X_i-x)/h)} \\ &+ \frac{\sum_{i=1}^n \rho(X_i,x)K((X_i-x)/h)}{\sum_{i=1}^n K((X_i-x)/h)} + \frac{\sum_{i=1}^n \epsilon_i K((X_i-x)/h)}{\sum_{i=1}^n K((X_i-x)/h)} \\ &:= F(x) + (I) + (II) + (III),\end{aligned}$$

where we have used the fact that the sums are invariant w.r.t. the order of summation and thus can be expressed in terms of the original i.i.d. variables (X_i). By the triangle inequality, non-negativity of the kernel, and lemma 3.2,

$$\begin{aligned}|(II)| &\leq \frac{\sum_{i=1}^n |\rho(X_i,x)|K((X_i-x)/h)}{\sum_{i=1}^n K((X_i-x)/h)} \\ &\leq L\frac{\sum_{i=1}^n |X_i-x)|^\beta K((X_i-x)/h)\mathbb{1}_{((X_i-x)/h)<1}}{\sum_{i=1}^n K((X_i-x)/h)} \\ &\leq Lh^\beta\end{aligned}$$

Similarly, (6) entails

$$|(III)| \leq \max_{1\leq i\leq n} |\epsilon_i| = O_p(n^{-1/2}).$$

(I) writes

$$(I) = \sum_{m=1}^{\ell} \frac{F^{(m)}(x)}{m!} \frac{s_m(x,h)}{s_0(x,h}.$$

Proposition (3.3) yields

$$s_m(x,h) = \begin{cases} f(x)\mu_m(K)h^m + o(h^m) + O_p\left(\frac{h^m}{\sqrt{hn}}\right) & \text{if } m \text{ is even,} \\ f'(x)\mu_{m+1}(K)h^{m+1} + o(h^{m+1}) + O_p\left(\frac{h^m}{\sqrt{hn}}\right) & \text{if } m \text{ is odd,} \end{cases}$$

Since $nh \to \infty$, the $O_P(h^m/\sqrt{hn})$ term is negligible w.r.t. h^m, for m even. Since K is a symmetric pdf with compact support, $\mu_0(K) = 1$, $\mu_1(K) = 0, \mu_2(K) > 0$. Hence, in (I) the terms with $m > 2$ are negligible. Therefore, at every point x s.t. $f(x) > 0$, one has, if $1 < \beta < 2$,

$$\begin{aligned}(I) = f(x)\frac{s_1(x,h)}{s_0(x,h)} &= f'(x)\mu_2(K)h^2 + o(h^2) + O_p\left(\frac{h}{\sqrt{hn}}\right) \\ &= f'(x)\mu_2(K)h^2 + o(h^2) + o_p\left(n^{-1/2}\right),\end{aligned}$$

so that $(I) = o(h^\beta) + o_p(n^{-1/2})$, and if $\beta \geq 2$,

$$\begin{aligned}(I) = f(x)\frac{s_1(x,h)}{s_0(x,h)} + \frac{f'(x)s_2(x,h)}{2s_0(x,h)} \\ = \frac{3f'(x)}{2}\mu_2(K)h^2 + o(h^2) + o_p\left(n^{-1/2}\right).\end{aligned}$$

Collecting all elements, we obtain, if $1 < \beta < 2$,

$$|\hat{F}_0(x) - F(x)| \leq Lh^\beta + o(h^\beta) + O_p(n^{-1/2}),$$

and, if $\beta \geq 2$,

$$\hat{F}_0(x) = F(x) + \frac{3f'(x)}{2}\mu_2(K)h^2 + o(h^2) + O_p(n^{-1/2}).$$

□

Proof of corollary 3.5. i) Any bandwidth of the form $h \sim n^{-\alpha}$, with $0 < \alpha < 1$, will yield a convergent estimator, i.e. s.t. $h \downarrow 0$ and $nh \to \infty$. If $\beta > 2$, then, for $\alpha > 1/4$, the term in h^2 becomes negligible compared to $n^{-1/2}$. If $1 < \beta < 2$, then, for $\alpha > 1/(2\beta)$, the term in $h\beta$ becomes negligible compared to $n^{-1/2}$. Since $\beta > 1$, $\alpha > 1/2$ suffices.

ii) Follows from i) by dominated convergence.

□

Proof of Corollary 3.6. Follows by inspection from the proof of Theorem 3.4. □

Acknowledgments

Olivier P. Faugeras acknowledges funding from ANR under grant ANR-17-EURE-0010 (Investissements d'Avenir program).

References

Abdous, B. and E. Bensaid (2001). "Multivariate local polynomial fitting for a probability distribution function and its partial derivatives". In: *J. Nonparametr. Statist.* 13(1), pp. 77–94. ISSN: 1048-5252. DOI: 10.1080/10485250008832843. URL: https://doi.org/10.1080/10485250008832843.

Aitken, Alexander C (1936). "IV.—On least squares and linear combination of observations". In: *Proceedings of the Royal Society of Edinburgh* 55, pp. 42–48.

Altman, N. S. (1990). "Kernel smoothing of data with correlated errors". In: *J. Amer. Statist. Assoc.* 85(411), pp. 749–759. ISSN: 0162-1459. URL: http://links.jstor.org/sici?sici=0162-1459(199009)85:411%3C749:KSODWC%3E2.0.CO;2-Q&origin=MSN.

Altomare, Francesco and Michele Campiti (1994). *Korovkin-type approximation theory and its applications*. Vol. 17. De Gruyter Studies in Mathematics. Appendix A by Michael Pannenberg and Appendix B by Ferdinand Beckhoff. Walter de Gruyter & Co., Berlin, pp. xii+627. ISBN: 3-11-014178-7. DOI: 10.1515/9783110884586. URL: https://doi.org/10.1515/9783110884586.

Anscombe, F. J. (1948). "The transformation of Poisson, binomial and negative-binomial data". In: *Biometrika* 35, pp. 246–254. ISSN: 0006-3444. DOI: 10.1093/biomet/35.3-4.246. URL: https://doi.org/10.1093/biomet/35.3-4.246.

Arnold, Barry C., Enrique Castillo, and José Marı'a Sarabia (2009a). "Multivariate order statistics via multivariate concomitants". In: *J. Multivariate Anal.* 100(5), pp. 946–951. ISSN: 0047-259X. DOI: 10.1016/j.jmva.2008.09.011. URL: https://doi.org/10.1016/j.jmva.2008.09.011.

Arnold, Barry C., Enrique Castillo, and José Marı'a Sarabia (2009b). "On multivariate order statistics. Applications to ranked set sampling". In: *Comput. Statist. Data Anal.* 53(12), pp. 4555–4569. ISSN: 0167-9473. DOI: 10.1016/j.csda.2009.05.011. URL: https://doi.org/10.1016/j.csda.2009.05.011.

Azzalini, A. (1981). "A note on the estimation of a distribution function and quantiles by a kernel method". In: *Biometrika* 68(1), pp. 326–328. ISSN: 0006-3444. DOI: 10.1093/biomet/68.1.326. URL: https://doi.org/10.1093/biomet/68.1.326.

Biau, Gérard and Luc Devroye (2015). *Lectures on the nearest neighbor method*. Springer Series in the Data Sciences. Springer, Cham, pp. ix+290. DOI: 10.1007/978-3-319-25388-6. URL: https://doi.org/10.1007/978-3-319-25388-6.

Bloomfield, Peter and William L. Steiger (1983). *Least absolute deviations*. Vol. 6. Progress in Probability and Statistics. Theory, applications, and algorithms. Birkhäuser Boston, Inc., Boston, MA, pp. xiv+349. ISBN: 0-8176-3157-7.

Bos, Thijs and Johannes Schmidt-Hieber (2024). "A supervised deep learning method for nonparametric density estimation". In: *Electron. J. Statist.* 18(2), pp. 5601–5658. ISSN: 1935-7524. DOI: 10.1214/24-EJS2332.

Chacón, José E., Pablo Monfort, and Carlos Tenreiro (2014). "Fourier methods for smooth distribution function estimation". In: *Statist. Probab. Lett.* 84, pp. 223–230. ISSN: 0167-7152. DOI: 10.1016/j.spl.2013.10.010. URL: https://doi.org/10.1016/j.spl.2013.10.010.

Chacón, José E. and Alberto Rodríguez-Casal (2010). "A note on the universal consistency of the kernel distribution function estimator". In: *Statistics and Probability Letters* 80(17), pp. 1414–1419. ISSN: 0167-7152. DOI: https://doi.org/

10.1016/j.spl.2010.05.007. URL: https://www.sciencedirect.com/science/article/pii/S0167715210001434.

Chernozhukov, Victor et al. (2017). "Monge-Kantorovich depth, quantiles, ranks and signs". In: *Ann. Statist.* 45(1), pp. 223–256. ISSN: 0090-5364. DOI: 10.1214/16-AOS1450. URL: https://doi.org/10.1214/16-AOS1450.

Chung, Kai-Lai (1949). "An estimate concerning the Kolmogoroff limit distribution". In: *Trans. Amer. Math. Soc.* 67, pp. 36–50. ISSN: 0002-9947. DOI: 10.2307/1990415. URL: https://doi.org/10.2307/1990415.

David, H. A. and H. N. Nagaraja (2003). *Order statistics*. Third. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, pp. xvi+458. ISBN: 0-471-38926-9. DOI: 10.1002/0471722162. URL: https://doi.org/10.1002/0471722162.

Durbin, James and M. Knott (1972). "Components of Cramér-von Mises statistics. I". In: *J. Roy. Statist. Soc. Ser. B* 34, pp. 290–307. ISSN: 0035-9246. URL: http://links.jstor.org/sici?sici=0035-9246(1972)34:2%3C290:COCMSI%3E2.0.CO;2-5&origin=MSN.

Dvoretzky, A., J. Kiefer, and J. Wolfowitz (1956). "Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator". In: *Ann. Math. Statist.* 27, pp. 642–669. ISSN: 0003-4851. DOI: 10.1214/aoms/1177728174. URL: https://doi.org/10.1214/aoms/1177728174.

Eilers, Paul H. C. and Brian D. Marx (1996). "Flexible smoothing with B-splines and penalties". In: *Statist. Sci.* 11(2). With comments and a rejoinder by the authors, pp. 89–121. ISSN: 0883-4237,2168-8745. DOI: 10.1214/ss/1038425655. URL: https://doi.org/10.1214/ss/1038425655.

Eilers, Paul HC and Brian D Marx (2010). "Splines, knots, and penalties". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 2(6), pp. 637–653.

Eilers, Paul HC and Brian D Marx (2021). *Practical smoothing: The joys of P-splines*. Cambridge University Press.

Fan, J. and I. Gijbels (1996). *Local polynomial modelling and its applications*. Vol. 66. Monographs on Statistics and Applied Probability. Chapman & Hall, London, pp. xvi+341. ISBN: 0-412-98321-4.

Faugeras, Olivier P. (2026). "The Stick-Breaking and Ordering Representation of Compositional Data: Copulas and Regression models". In: *Austrian Journal of Statistics* 55(1), pp. 1–31. URL: https://hal.science/hal-04409705.

Fernholz, Luisa Turrin (1991). "Almost sure convergence of smoothed empirical distribution functions". In: *Scand. J. Statist.* 18(3), pp. 255–262. ISSN: 0303-6898.

Finkelstein, Helen (1971). "The law of the iterated logarithm for empirical distributions". In: *Ann. Math. Statist.* 42, pp. 607–615. ISSN: 0003-4851. DOI: 10.1214/aoms/1177693410. URL: https://doi.org/10.1214/aoms/1177693410.

Frasso, Gianluca and Paul H. C. Eilers (2015). "L- and V-curves for optimal smoothing". In: *Stat. Model.* 15(1), pp. 91–111. ISSN: 1471-082X,1477-0342. DOI: 10.1177/1471082X14549288. URL: https://doi.org/10.1177/1471082X14549288.

Genon-Catalot, Valentine, Catherine Laredo, and Michael Nussbaum (2002). "Asymptotic Equivalence of Estimating a Poisson Intensity and a Positive Diffusion Drift". In: *The Annals of Statistics* 30(3), pp. 731–753. DOI: 10.1214/aos/1028674838.

Hampel, Frank R. et al. (1986). *Robust statistics*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. The approach based on influence functions. John Wiley & Sons, Inc., New York, pp. xxiv+502. ISBN: 0-471-82921-8.

Hart, Jeffrey D (1991). "Kernel regression estimation with time series errors". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 53(1), pp. 173–187.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The elements of statistical learning*. Second. Springer Series in Statistics. Data mining, inference, and prediction. Springer, New York, pp. xxii+745. ISBN: 978-0-387-84857-0. DOI: 10.1007/978-0-387-84858-7. URL: https://doi.org/10.1007/978-0-387-84858-7.

Hjort, N. L. and M. C. Jones (1996). "Locally parametric nonparametric density estimation". In: *Ann. Statist.* 24(4), pp. 1619–1647. ISSN: 0090-5364. DOI: 10.1214/aos/1032298288. URL: https://doi.org/10.1214/aos/1032298288.

Ibragimov, I. A. and R. Z. Khasminskii (1981). *Statistical estimation*. Vol. 16. Applications of Mathematics. Asymptotic theory, Translated from the Russian by Samuel Kotz. Springer-Verlag, New York-Berlin, pp. vii+403. ISBN: 0-387-90523-5.

Kariya, Takeaki and Hiroshi Kurata (2004). *Generalized least squares*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester, pp. xiv+289. ISBN: 0-470-86697-7. DOI: 10.1002/0470866993. URL: https://doi.org/10.1002/0470866993.

Korovkin, P. P. (1953). "On convergence of linear positive operators in the space of continuous functions". In: *Doklady Akad. Nauk SSSR (N.S.)* 90, pp. 961–964.

Kuhn and Max (2008). "Building Predictive Models in R Using the caret Package". In: *Journal of Statistical Software* 28(5), pp. 1–26. DOI: 10.18637/jss.v028.i05. URL: https://www.jstatsoft.org/index.php/jss/article/view/v028i05.

Lejeune, Michel and Pascal Sarda (1992). "Smooth estimators of distribution and density functions". In: *Comput. Statist. Data Anal.* 14(4), pp. 457–471. ISSN: 0167-9473. DOI: 10.1016/0167-9473(92)90061-J. URL: https://doi.org/10.1016/0167-9473(92)90061-J.

Lindsey, J. K. (1974a). "Comparison of probability distributions". In: *J. Roy. Statist. Soc. Ser. B* 36, pp. 38–47. ISSN: 0035-9246. URL: http://links.jstor.org/sici?sici=0035-9246(1974)36:1%3C38:COPD%3E2.0.CO;2-9&origin=MSN.

Lindsey, J. K. (1974b). "Construction and comparison of statistical models". In: *J. Roy. Statist. Soc. Ser. B* 36, pp. 418–425. ISSN: 0035-9246. URL: http://links.jstor.org/sici?sici=0035-9246(1974)36:3%3C418:CACOSM%3E2.0.CO;2-Y&origin=MSN.

Loader, Clive (1999). *Local regression and likelihood*. Statistics and Computing. Springer-Verlag, New York, pp. xiv+290. ISBN: 0-387-98775-4.

Loader, Clive R. (1996). "Local likelihood density estimation". In: *Ann. Statist.* 24(4), pp. 1602–1618. ISSN: 0090-5364. DOI: 10.1214/aos/1032298287. URL: https://doi.org/10.1214/aos/1032298287.

Maronna, Ricardo A., R. Douglas Martin, and Victor J. Yohai (2006). *Robust statistics*. Wiley Series in Probability and Statistics. Theory and methods. John

Wiley & Sons, Ltd., Chichester, pp. xx+403. DOI: 10.1002/0470010940. URL: https://doi.org/10.1002/0470010940.

Massart, P. (1990). "The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality". In: *Ann. Probab.* 18(3), pp. 1269–1283. ISSN: 0091-1798. URL: http://links.jstor.org/sici?sici=0091-1798(199007)18:3%3C1269:TTCITD%3E2.0.CO;2-Q&origin=MSN.

Mosler, Karl (2013). "Depth statistics". In: *Robustness and complex data structures*. Springer, Heidelberg, pp. 17–34. DOI: 10.1007/978-3-642-35494-6_2. URL: https://doi.org/10.1007/978-3-642-35494-6_2.

Nadaraya, È. A. (1964a). "Some new estimates for distribution functions". In: *Teor. Verojatnost. i Primenen.* 9, pp. 550–554. ISSN: 0040-361x.

Nadaraya, Elizbar A (1964b). "On estimating regression". In: *Theory of Probability & Its Applications* 9(1), pp. 141–142.

Nussbaum, Michael (1996). "Asymptotic equivalence of density estimation and Gaussian white noise". In: *Ann. Statist.* 24(6), pp. 2399–2430. ISSN: 0090-5364. DOI: 10.1214/aos/1032181160. URL: https://doi.org/10.1214/aos/1032181160.

Opsomer, Jean, Yuedong Wang, and Yuhong Yang (2001). "Nonparametric Regressin with Correlated Errors". In: *Statistical Science* 16(2), pp. 134–153. DOI: 10.1214/ss/1009213287. URL: https://doi.org/10.1214/ss/1009213287.

Payton, M. E., L. J. Young, and J. H. Young (1989). "Bounds for the difference between median and mean of beta and negative binomial distributions". In: *Metrika* 36(6), pp. 347–354. ISSN: 0026-1335. DOI: 10.1007/BF02614111. URL: https://doi.org/10.1007/BF02614111.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Ray, Kolyan and Johannes Schmidt-Hieber (2016). "The Le Cam Distance between Density Estimation, Poisson Processes and Gaussian White Noise". In: *Mathematical Statistics and Learning* 1(2), pp. 101–170. DOI: 10.4171/MSL/3.

Ray, Kolyan and Johannes Schmidt-Hieber (2019). "Asymptotic Nonequivalence of Density Estimation and Gaussian White Noise for Small Density". In: *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* 55(4), pp. 2195–2208. DOI: 10.1214/18-AIHP946.

Reiss, R.-D. (1989). *Approximate distributions of order statistics*. Springer Series in Statistics. With applications to nonparametric statistics. Springer-Verlag, New York, pp. xii+355. ISBN: 0-387-96851-2. DOI: 10.1007/978-1-4613-9620-8. URL: https://doi.org/10.1007/978-1-4613-9620-8.

Riazoshams, Hossein, Habshah Midi, and Gebrenegus Ghilagaber (2018). *Robust nonlinear regression: with applications using R*. John Wiley & Sons.

Rousseeuw, Peter J. and Annick M. Leroy (1987). *Robust regression and outlier detection*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, Inc., New York, pp. xvi+329. ISBN: 0-471-85233-3. DOI: 10.1002/0471725382. URL: https://doi.org/10.1002/0471725382.

Scott, David W. (2015). *Multivariate density estimation*. Second. Wiley Series in Probability and Statistics. Theory, practice, and visualization. John Wiley & Sons, Inc., Hoboken, NJ, pp. xviii+350. ISBN: 978-0-471-69755-8.

Servien, Rémi (2009). "Estimation de la fonction de répartition: revue bibliographique". In: *J. SFdS* 150(2), pp. 84–104.

Shorack, Galen R. (2000). *Probability for Statisticians*. Springer.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, pp. x+175. ISBN: 0-412-24620-1. DOI: 10.1007/978-1-4899-3324-9. URL: https://doi.org/10.1007/978-1-4899-3324-9.

Smirnoff, N. (1939). "Sur les écarts de la courbe de distribution empirique". In: *Rec. Math. N.S. [Mat. Sbornik]* 6(48), pp. 3–26.

Tsybakov, Alexandre B. (2004). *Introduction à l'estimation non-paramétrique*. Vol. 41. Mathématiques & Applications (Berlin) [Mathematics & Applications]. Springer-Verlag, Berlin, pp. x+175. ISBN: 3-540-40592-5.

Tukey, John W. (1975). "Mathematics and the picturing of data". In: *Proceedings of the International Congress of Mathematicians (Vancouver, B.C., 1974), Vol. 2*. Canad. Math. Congr., Montreal, QC, pp. 523–531.

Wand, M. P. and M. C. Jones (1995). *Kernel smoothing*. Vol. 60. Monographs on Statistics and Applied Probability. Chapman and Hall, Ltd., London, pp. xii+212. ISBN: 0-412-55270-1. DOI: 10.1007/978-1-4899-4493-1. URL: https://doi.org/10.1007/978-1-4899-4493-1.

Watson, Geoffrey S (1964). "Smooth regression analysis". In: *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 359–372.

Weller, Grant B and William F Eddy (2015). "Multivariate order statistics: Theory and application". In: *Annual Review of Statistics and Its Application* 2(1), pp. 237–257.

Winter, B. B. (1973). "Strong uniform consistency of integrals of density estimators". In: *Canad. J. Statist.* 1(2), pp. 247–253. ISSN: 0319-5724. DOI: 10.2307/3315003. URL: https://doi.org/10.2307/3315003.

Yamato, Hajime (1972/73a). "Some statistical properties of estimators of density and distribution functions". In: *Bull. Math. Statist.* 15(1-2), pp. 113–131. ISSN: 0007-4993.

Yamato, Hajime (1972/73b). "Uniform convergence of an estimator of a distribution function". In: *Bull. Math. Statist.* 15(3-4), pp. 69–78.

Zuo, Yijun and Robert Serfling (2000). "General notions of statistical depth function". In: *Ann. Statist.* 28(2), pp. 461–482. ISSN: 0090-5364. DOI: 10.1214/aos/1016218226. URL: https://doi.org/10.1214/aos/1016218226.