Iterated Bernstein operators for distribution function and density estimation: Balancing between the number of iterations and the polynomial degree

Claude Manté

Iterated Bernstein operators for distribution function and density estimation: balancing between the number of iterations and the polynomial degree.

Claude Manté[a]

[a] *Aix-Marseille Université, Université du Sud Toulon-Var, CNRS/INSU, IRD, MIO, UM 110, Campus de Luminy, Case 901, F13288 Marseille Cedex 09, France.*
tel: (+33) 486 090 631 fax: (+33) 486 090 641

Abstract

Despite its slow convergence, the use of the Bernstein polynomial approximation is becoming more frequent in Statistics, especially for density estimation of compactly supported probability distributions. This is due to its numerous attractive properties, from both an approximation (uniform shape-preserving approximation, *etc.*) and a statistical (*bona fide* estimation, low boundary bias, *etc.*) point of view. An original method for estimating distribution functions and densities with Bernstein polynomials is proposed, which takes advantage of results about the eigenstructure of the Bernstein operator to refine a convergence acceleration method. Furthermore, an original adaptative method for choosing the degree of the polynomial is worked out. The method is successfully applied to two data-sets which are important benchmarks in the field of Density Estimation.

Keywords: Density Estimation, Bernstein operator, roots of operators, regular histogram, shape restriction, Gnedenko test

Email address: claude.mante@mio.osupytheas.fr (Claude Manté)

1. Introduction

Although S. Bernstein simultaneously introduced both the polynomials and the operator that bear his name in his famous constructive proof of the Stone-Weierstrass theorem (Bernstein , 1912), both of these objects naturally split up with time. While there is a large interest in the Bernstein operator in the literature on Approximation Theory, (see for instance Cooper and Waldron (2000); Sevy (1993, 1995); Sahai (2004)), researchers from other disciplines essentially focus on Bernstein polynomials. For instance, the attractive properties of this approximation prompted statisticians to apply it to Density Estimation (Vitale , 1975; Babu et al. , 2002; Bouezmarni and Rolin , 2007; Leblanc , 2010, 2012a,b), Regression (A. Marco and J.J. Martinez , 2010; Curtis and Ghosh , 2011; Wang and Ghosh , 2012) or Bayesian Inference (Petrone , 1999). However, most of these authors paid little attention to the Bernstein operator itself.

Nevertheless, an operator is attached to a pair of vector spaces, and not to particular bases of these spaces. We highlight in Section 3 that Bernstein polynomials consist in a natural **output** basis for the eponym operator, while the natural **input** basis is a Lagrange polynomial basis (see also Cooper and Waldron (2000), Section 5). In addition, we must take into account the pair of bases associated with the eigendecompositon of the Bernstein operator, given by Cooper and Waldron (2000). Bearing in mind a generalization of the Sevy convergence acceleration method (Sevy , 1993, 1995), we further investigate in Section 3 the matrix representation of powers of the Bernstein operator with respect to these bases. This enables us to define first, in

Section 4, fractional Bernstein operators and second, in Section 5, fractional Sevy approximation sequences. This constitutes the basis for refining results obtained by Manté (2012), where both the distribution function and the density approximation were obtained using Sevy's iteration scheme.

Now, roughly speaking, density estimation or approximation by Bernstein polynomials (Babu et al. , 2002; Bouezmarni and Rolin , 2007; Leblanc , 2010, 2012a,b; Manté , 2012) consists in fitting a Bernstein polynomial of some order m on a distribution function, and in differentiating it. More precisely, these authors estimate the distribution function (d.f.) F associated with a random variable X from m values of the empirical distribution function (e.d.f.) F_N obtained from a N-sample of X :

$$\widetilde{F}_{N,m}(x) := \sum_{k=0}^{m} F_N(\frac{k}{m})\, w_{m,k}(x)\,,$$

where $w_{n,j}(x) := \binom{n}{j} x^j\,(1-x)^{n-j}$. The choice of an optimal number of bins m^* is always a critical step. In the density estimation setting, most authors recommend either choosing $m^* = \nu(N)$, where N is the sample size and ν is some function stemming from asymptotic results (Babu et al. , 2002; Leblanc , 2010, 2012a), or else obtaining m^* from cross-validation methods (Bouezmarni and Rolin , 2007; Leblanc , 2010).

In Section 6 we propose another method, starting with the Babu et al. (2002) upper value $m_0 := N/\ln(N)$. It consists in selecting $m^* \leq m_0$ in order that the same optimal m^* should be obtained with a high probability from different N-samples (stability), and that the "coarsened" distribution functions associated with these m^* bins should be close to the classical empirical distribution function F_N (fidelity). The method is tested on real data

in Section 7.

2. Notation

We will work in the Banach space $C[0,1]$ of continuous functions on $[0,1]$, equipped with the Chebyshev norm $\|f\| := \max_{x\in[0,1]} |f(x)|$. $\mathfrak{P}_n$ denotes the subspace of $C[0,1]$ consisting of polynomials of degree $k \leq n$, and $\overline{\mathfrak{P}_n}$ denotes the complement of $\mathfrak{P}_1$ in $\mathfrak{P}_n$ *i.e.* the vector space of polynomials of degree $1 < k \leq n$.

Consider an operator $U : C[0,1] \to C[0,1]$; for $n \geq 2$ (fixed), its restriction to $\mathfrak{P}_n$ (*i.e.* the operator $U|_{\mathfrak{P}_n} : \mathfrak{P}_n \to C[0,1]$ such that $\forall f \in \mathfrak{P}_n$, $U|_{\mathfrak{P}_n}(f) = U(f)$) will be denoted $\overset{\circ}{U}$, and its restriction to $\overline{\mathfrak{P}_n}$ will be denoted $\overline{U}$. For the sake of simplicity, the restrictions of the identity operator to these subspaces will be denoted 1, instead of $\overset{\circ}{1}$ or $\overline{1}$.

In the finite dimensional setting, we will use the matrix p-norm (or ℓ^p-norm) $\|U\|_p := \sup_{v\neq 0} \frac{\|U(v)\|_p}{\|v\|_p}$ where $\|v\|_p$ is the usual vector ℓ^p-norm. Notice that $\|U\|_1$ and $\|U\|_\infty$ are the greatest sum of the absolute values of the matrix elements along columns and rows, respectively, while $\|U\|_2$ is the spectral norm (Farouki , 1991). In this setting, $Mat(U; L_n, W_n)$ will denote the matrix representation of the operator U with respect to the bases L_n and W_n.

Finally, the expression $Y \overset{\mathfrak{L}}{=} X$ denotes that both of the random variables X and Y obey the same probability law. The integer value of some real number x will be denoted $\lfloor x \rfloor$.

3. Expression of powers of the Bernstein operator into different bases

The Bernstein operator $B_n : C[0,1] \to C[0,1]$ is defined (Cooper and Waldron , 2000; Manté , 2012; Sevy , 1995) by:

$$B_n[f](x) := \sum_{j=0}^{n} w_{n,j}(x) f(\frac{j}{n}),$$

with $w_{n,j}(x) := \binom{n}{j} x^j (1-x)^{n-j}$. Of course, its image $\mathcal{R}(B_n)$ is included in $\mathfrak{P}_n$. In this section, we will focus on the matrix representation of powers of B_n with respect to three bases of $\mathfrak{P}_n$: Lagrange and Bernstein bases, and the eigenfunctions of B_n.

3.1. Expression of powers of B_n relative to Lagrange and Bernstein bases

First, let us consider the Lagrange interpolation operator $\mathcal{L}_n : C[0,1] \to C[0,1]$, defined by

$$\mathcal{L}_n[f](x) := \sum_{j=0}^{n} \ell_{n,j}(x) f(\frac{j}{n}),$$

where $\ell_{n,j}(x) := \prod_{\substack{k=0 \\ k \neq j}}^{n} \frac{n\,x-k}{j-k}$ is the jth Lagrange polynomial in the equally spaced case. Clearly, $\mathcal{R}(\mathcal{L}_n) = \mathfrak{P}_n$ and, since $\mathcal{L}_n$ is idempotent and the Lebesgue constant $\|\mathcal{L}_n\| = \max_{\|f\| \neq 0} \frac{\|\mathcal{L}_n[f]\|}{\|f\|} \sim \frac{2^n}{e\,n\,\log(n)}$ (see Mills and Smith (1992)) is bounded for any finite n, $\mathcal{L}_n$ is the projection onto $\mathfrak{P}_n$. Consequently, any $f \in C[0,1]$ is the direct sum of two components: $\mathcal{L}_n[f]$ and the " Lagrange residual" $(f - \mathcal{L}_n[f])$.

Lemma 1. $\forall k \geq 1,\ B_n^k = \overset{\circ}{B}{}_n^k \circ \mathcal{L}_n$, *where* $\overset{\circ}{B}{}_n^k := \left(\overset{\circ}{B}_n\right)^k$ *denotes the power of order* k *of the restricted operator.*

Proof. Because $\mathcal{L}_n$ is interpolatory, we can write:

$$B_n : C[0,1] \xrightarrow{\mathcal{L}_n} \mathfrak{P}_n \xrightarrow{\overset{\circ}{B}_n} \mathfrak{P}_n.$$

In other words, $B_n = \overset{\circ}{B}_n \circ \mathcal{L}_n$; furthermore, since $\mathcal{L}_n$ is the projection onto $\mathfrak{P}_n$, $\forall k \geq 1,\ \mathcal{L}_n \circ \overset{\circ}{B}{}_n^k = \overset{\circ}{B}{}_n^k$ □

Consider now a polynomial $P \in \mathfrak{P}_n$; we have on the one hand $\mathcal{L}_n[P](x) = \sum_{j=0}^{n} \ell_{n,j}(x)\, P(\frac{j}{n})$ and on the other hand $\overset{\circ}{B}_n[P](x) = \sum_{j=0}^{n} w_{n,j}(x) P(\frac{j}{n})$. Thus, with respect to the bases $L_n := \{\ell_{n,j}(x), 0 \leq j \leq n\}$ and $W_n := \{w_{n,j}(x), 0 \leq j \leq n\}$ the matrix of $\overset{\circ}{B}_n$ is the identity matrix: $Mat\left(\overset{\circ}{B}_n; L_n, W_n\right) = I_n$. Let us denote $LW_{[n]}$ the transformation matrix associated with the bases L_n and W_n , whose jth column consists in the coordinates of $w_{n,j}$ in the basis L_n.

Lemma 2. *The matrix of* $\overset{\circ}{B}{}_n^k$ *with respect to the bases* L_n *and* W_n *is* $Mat\left(\overset{\circ}{B}{}_n^k; L_n, W_n\right) = LW_{[n]}^{k-1}$.

Proof. One can easily verify that $LW_{[n]\,i,j} = w_{n,j}\left(\frac{i}{n}\right)$; consequently, $Mat\left(\overset{\circ}{B}_n; W_n, W_n\right) = LW_{[n]}$. Thus, the iterated operator of order k can be represented by the diagram:

$$B_n^k : C[0,1] \xrightarrow{\mathcal{L}_n} (\mathfrak{P}_n, L_n) \xrightarrow{I_n} (\mathfrak{P}_n, W_n) \overset{LW_{[n]}^{k-1}}{\longrightarrow} (\mathfrak{P}_n, W_n)$$

and $Mat\left(\overset{\circ}{B}{}_n^k; L_n, W_n\right) = LW_{[n]}^{k-1}$ □

3.2. Expression of powers of B_n relative to the eigenfunctions of B_n

At present, the focus is on the eigenstructure of B_n, which was completely elucidated by Cooper and Waldron (2000). They demonstrated the following theorem.

Theorem 1. *The Bernstein operator can be represented in the diagonal form*

$$B_n[f] = \sum_{j=0}^{n} \lambda_j^{[n]} \pi_j^{[n]} \mu_j^{[n]}(f), \tag{1}$$

where $f \in C[0,1]$, $\lambda_j^{[n]}$ and $\pi_j^{[n]}$ are the eigenvalues and eigenfunctions of B_n, and $\mu_j^{[n]}$ are the dual functionals to $\pi_j^{[n]}$.

The eigenvalues are given by $\lambda_j^{[n]} = \frac{n!}{(n-j)!\, n^j}$, while $\pi_j^{[n]}$ is a polynomial of degree j , which can be calculated with a recurrence formula given in Cooper and Waldron (2000). As for the $\mu_j^{[n]}$, they constitute a basis for the dual space $\mathfrak{P}_n^* \subsetneq C[0,1]^*$, such that $\left\langle \mu_j^{[n]}, \pi_k^{[n]} \right\rangle = \delta_{j,k}\, \forall j,k$.

Corollary 1. *Using the classical notation $u \otimes v^*(w) := u \langle v^*, w \rangle$ (Bowen and Wang , 1976), we can rewrite Eq. (1) in an alternative form:*

$$B_n[f] = \sum_{j=0}^{n} \lambda_j^{[n]} \pi_j^{[n]} \otimes \pi_j^{*[n]} (\mathcal{L}_n[f]). \tag{2}$$

Proof: see the appendix.

Thus, B_n and $\overset{\circ}{B}_n$ have exactly the same eigenstructure. Denoting $\Lambda_{[n]}$ the diagonal matrix associated with the $\lambda_j^{[n]}$, we can now write the diagram:

$$B_n^k : C[0,1] \xrightarrow{\mathcal{L}_n} (\mathfrak{P}_n, L_n) \xrightarrow{L\Pi_{[n]}} \left(\mathfrak{P}_n, \Pi_{[n]}\right) \xrightarrow{\Lambda_{[n]}^k} \left(\mathfrak{P}_n, \Pi_{[n]}\right) \xrightarrow{\Pi W_{[n]}} (\mathfrak{P}_n, W_n) \tag{3}$$

where $L\Pi_{[n]}$ and $\Pi W_{[n]}$ are the transformation matrices associated with these bases.

4. Fractional Bernstein operators

We propose in this section, for any integer $K \geq 2$, a definition of the K^{th} "root" of the operator $G_n := (1 - B_n)$, denoted $G_n^{1/K}$; this will enable us to generalize iterated boolean sums of operators studied by Sevy (1993).

For a fixed continuous function f, consider the decomposition: $(1 - B_n)[f] = (f - \mathcal{L}_n[f]) + (\mathcal{L}_n[f] - B_n[f])$. While it is straightforward to write that $(\mathcal{L}_n[f] - B_n[f]) = \left(1 - \overset{\circ}{B}_n\right)[\mathcal{L}_n[f]]$, we do not have much information about the residual $f - \mathcal{L}_n[f]$. In fact, the only thing that can be said is that $\|f - \mathcal{L}_n[f]\| \leq (1 + \|\mathcal{L}_n\|)\|f\| \sim \frac{2^n}{e\, n \log(n)}\|f\|$ (Laurent , 1972; Mills and Smith , 1992). This does not matter here, because the objective is to compute expressions like $\left(1 - (1 - B_n)^I\right)[f]$ (see Section 5), and we have the following result (see appendix for proof).

Lemma 3. *For any integer I,* $\left(1 - (1 - B_n)^I\right) = \left(1 - \left(1 - \overset{\circ}{B}_n\right)^I\right) \circ \mathcal{L}_n$.

Consequently, we can proceed as if $f \in \mathcal{R}(\mathcal{L}_n)$, and we do not need to worry about the Lagrange residual. Now, thanks to Eq. (2), we have:

$$\overset{\circ}{G}_n \circ \mathcal{L}_n[f] := \left(1 - \overset{\circ}{B}_n\right)[\mathcal{L}_n[f]] = \sum_{j=0}^{n} \left(1 - \lambda_j^{[n]}\right) \pi_j^{[n]} \otimes \pi_j^{*[n]}(\mathcal{L}_n[f]).$$

Thus, $\overset{\circ}{G}_n$ can be considered as a symmetrical bilinear application, and its matrix relative to some basis $E_{[n]}$ of $\mathfrak{P}_n$ is $Mat\left(\overset{\circ}{G}_n; E_{[n]}, E_{[n]}\right) = Q_{[n]}\, \Gamma_{[n]}\, Q_{[n]}^t$, where $Q_{[n]}$ is orthogonal and $\Gamma_{[n]}$ is the diagonal matrix associated with the vector $(0, 0, 1/n, (3n-2)/n^2, \cdots, 1 - n!/n^n)$.

Consider now the restriction $\overline{B_n}$ of $\overset{\circ}{B}_n$ to $\overline{\mathfrak{P}_n}$. Since B_n reproduces only the linear polynomials (even quadratic polynomials are not reproduced by B_n

- see Walz (2000)), the operator $\overline{G_n} := \overline{1 - B_n}$ is injective. All its eigenvalues are positive, and the maximal one is $1 - \frac{n!}{n^n} < 1 - \sqrt{2\pi n}\, \exp\left(-n + \frac{1}{12n} - \frac{1}{360n^3}\right)$ (see Impens (2003)).

Since the maximum eigenvalue of $\overline{B_n}$ is $1 - \frac{1}{n}$, it is possible to define a new operator $\overline{G_n}^{(\alpha)}$ from the classical results below (valid in a much larger setting than ours).

Proposition 1. *(Kato (1995), Ch. 9&10)*

(1) Let T be an operator of finite trace (trace class) in a separable Hilbert space $\mathcal{H}$, such that its spectral radius is smaller than 1. Then we may define the operator

$$\log(1+T) := \sum_{k=1}^{\infty} (-1)^{k-1} \frac{T^k}{k},$$

which also belongs to the trace class.

(2) Let T be a bounded operator defined on a Banach space. Consider the Taylor series:

$$\sum_{k=0}^{\infty} \frac{(-1)^k u^k T^k}{k!}.$$

It is absolutely convergent for any complex number u, and defines an operator denoted $\exp(-uT)$.

Using the first part of the Proposition above, we can first define the operator $\log\left(\overline{G_n}\right)$ and afterwards, thanks to the second part, we can define for any $\alpha > 0$ the operator we need:

Definition 1.

$$\overline{G_n}^{(\alpha)} := \exp\left(\alpha \log\left(\overline{G_n}\right)\right). \tag{4}$$

The matrix representation of this new operator is simple; it is given in the following result (see the appendix for a proof).

Proposition 2. *Suppose* $\overline{E_{[n]}}$ *is some basis of* $\overline{\mathfrak{P}_n}$. *Then,* $Mat\left(\overline{G_n}^{(\alpha)}; \overline{E_{[n]}}, \overline{E_{[n]}}\right) = \overline{Q_{[n]}}\ \overline{\Gamma_{[n]}}^{(\alpha)}\ \overline{Q_{[n]}}^t$, *where* $\overline{\Gamma_{[n]}}^{(\alpha)}$ *is the diagonal matrix associated with the vector* $\left(\left(\frac{1}{n}\right)^\alpha, \left(\frac{3n-2}{n^2}\right)^\alpha, \cdots, \left(1-\frac{n!}{n^n}\right)^\alpha\right)$ *and* $\overline{Q_{[n]}}$ *is orthogonal.*

5. Interpolating Sevy sequences

In order to accelerate the convergence of Bernstein approximations, Sevy (1993, 1995) proposed to replace B_n by the iterated operator

$$\mathfrak{I}_n^I := \left(1-(1-B_n)^I\right). \tag{5}$$

This method was re-discovered by Sahai (2004), who noticed that one can write $C^0[0,1] \ni F = B_n[F]+E$, where $E \in C^0[0,1]$ is an unknown " Bernstein residual" which can be approximated by $B_n[E]$. Then, $B_n[F] + B_n[E] = \left(1-(1-B_n)^2\right)[F]$ is a better approximation of F than $B_n[F]$, and so on... Sevy proved the following result :

Theorem 2. *(Sevy (1995), see also Cooper and Waldron (2000)) For some fixed* $n \geq 1$ *and any function* F *defined on [0,1],*

$$\left\|\mathfrak{I}_n^I[F] - \mathcal{L}_n[F]\right\| \underset{I\to\infty}{\longrightarrow} 0\,. \tag{6}$$

Thus, Sevy sequences build a bridge between Bernstein approximation (which has nice shape-preserving properties, but converges slowly) and Lagrange interpolation, which is notoriously a bad approximate, especially in the case of equispaced knots (de Boor (1978, Ch. 2); see also Laurent (1972,

Ch. 5)). Both of these polynomials can have bad properties: the first one can be suspected of excessive smoothness (especially when the sample size is small or moderate), while the second one is generally "bumpy". Searching for a trade-off, Cooper and Waldron (2000) proposed to run across the whole segment $t\,B_n[F] + (1-t)\,\mathcal{L}_n[F]$, $0 \leq t \leq 1$. We will follow a different line, specific to density approximation, to work out another trade-off between both types of approximations.

Proposition 3. *Let $P \in \mathfrak{P}_n = \mathfrak{P}_1 \oplus \overline{\mathfrak{P}_n}$, and consider the associated decomposition: $P = P_1 + \bar{P}$. We have:*

$$\forall\, k \geq 1,\, \mathfrak{I}_n^k(P) = P_1 + \mathfrak{I}_n^k\left(\bar{P}\right).$$

Proof: see the appendix.

Because of Lemma 3 and the proposition above, $\forall\, f \in C[0,1]$, $\mathfrak{I}_n^k(f) = \mathcal{L}_1[f] + \mathfrak{I}_n^k(\mathcal{L}_n[f] - \mathcal{L}_1[f])$, with $\mathcal{L}_1[f](x) = x\,f(1) + (1-x)\,f(0)$ $\forall\, x \in [0,1]$. Consequently, it is quite natural to propose the following definition of fractional sequences.

Definition 2. *Let $K \geq 2$ be an integer, and $f \in C[0,1]$. The K-fractional Sevy approximation sequence of f is defined by:*

$$\mathfrak{I}_{n;K}^j[f] := \mathcal{L}_1[f] + \left(1 - \overline{G_n}^{(j/K)}\right)(\mathcal{L}_n[f] - \mathcal{L}_1[f]),\; j \geq 1.$$

This sequence interpolates Sevy's one, since $\mathfrak{I}_{n;K}^{j\,K}[f] = \mathfrak{I}_n^j(f)$.

Proposition 4. *The matrix of the restricted fractional operator is: $Mat\left(\overset{\circ j}{\mathfrak{I}}_{n;K}; L_n, W_n\right) = \Pi W_{[n]} \circ \Lambda_{[n]}^{(j/K)} \circ L\Pi_{[n]}$, where $\Lambda_{[n]}^{(j/K)}$ is the diagonal matrix associated with the vector $\left(1, 1, 1 - \left(\frac{1}{n}\right)^{(j/K)}, 1 - \left(\frac{3n-2}{n^2}\right)^{(j/K)}, \cdots, 1 - \left(1 - \frac{n!}{n^n}\right)^{(j/K)}\right)$.*

Proof. Using the blocks structure associated with the decomposition $\mathfrak{P}_n = \mathfrak{P}_1 \oplus \overline{\mathfrak{P}_n}$, we can see that $Mat\left(\overset{\circ}{\mathfrak{I}}{}^{j}_{n;K}; \Pi_{[n]}, \Pi_{[n]}\right) = \Lambda^{(j/K)}_{[n]}$. Thus, the fractional operator can be represented by a diagram similar to (3):

$$\mathfrak{I}^{j}_{n;K} : C[0,1] \xrightarrow{\mathcal{L}_n} (\mathfrak{P}_n, L_n) \xrightarrow{L\Pi_{[n]}} \left(\mathfrak{P}_n, \Pi_{[n]}\right) \xrightarrow{\Lambda^{(j/K)}_{[n]}} \left(\mathfrak{P}_n, \Pi_{[n]}\right) \xrightarrow{\Pi W_{[n]}} (\mathfrak{P}_n, W_n)$$

□

5.1. Numerical difficulties

Because of Lemmas 2 and 3, computing a classical Sevy sequence amounts to computing powers of the transformation matrix $LW_{[n]}$. Since $Mat\left(\overset{\circ}{B}_n; W_n, W_n\right) = LW_{[n]}$, the condition number of this matrix in the ℓ^2-norm is (Farouki , 1991): $\frac{\left\|LW_{[n]}\right\|_2}{\left\|LW^{-1}_{[n]}\right\|_2} = \frac{\lambda^{[n]}_0}{\lambda^{[n]}_n} = \frac{n^n}{n!} \approx \frac{e^n}{\sqrt{2\pi n}}$ (Cooper and Waldron , 2000; Impens , 2003). Thus, $LW_{[n]}$ is ill-conditioned in the ℓ^2-norm sense, and one must expect to encounter numerical problems when n is big enough. The situation is more complicated in the case of fractional sequences, since Proposition 4 shows that the matrix of the restricted operator depends on both of the transformation matrices $L\Pi_{[n]}$ and $\Pi W_{[n]}$. To our knowledge, the transformations between Lagrange polynomials, Bernstein polynomials, and the Bernstein operator eigenfunctions system have not been studied yet. However, it is well-known that the transformations between power and Bernstein bases (Farouki , 1991, 2012) or between Hermite and Bernstein bases (Hermann , 1996) are ill-conditioned.

The idea here is merely to control numerical errors in the computation of $\mathfrak{I}^{j}_{n;K}[f]$. First, notice that $Mat\left(\overset{\circ}{B}_n; L_n, W_n\right) = \Pi W_{[n]} \circ \Lambda_{[n]} \circ L\Pi_{[n]}$; thus, since $Mat\left(\overset{\circ}{B}_n; L_n, W_n\right) = I_n$, the matrix norms $\left\|\Pi W_{[n]} \circ \Lambda_{[n]} \circ L\Pi_{[n]} - I_n\right\|_1$

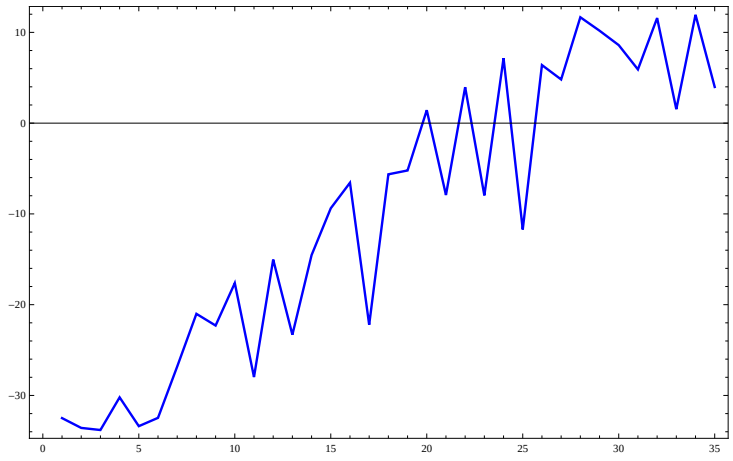

Figure 1:

and $\left\|\Pi W_{[n]} \circ \Lambda_{[n]} \circ L\Pi_{[n]} - I_n\right\|_\infty$ are convenient indicators of loss of numerical accuracy due to the ill-conditioning of the transformation matrices. In Figure 1, we plotted the logarithm of the second indicator for n ranging from 1 to 35 (a similar graph has been obtained for the first indicator). The reader can see that indeed the dimension of $\mathfrak{P}_n$ should not exceed $n = 21$. Furthermore, we must add to this difficulty the computational cost of the eigenfunctions which becomes prohibitive for $n \geq 22$. To sum up, for practical reasons, it seems necessary to restrict ourselves to polynomials of degree lower than 21.

6. Application of fractional sequences to distribution function and density estimation

Suppose F is a differentiable d.f. associated with a random variable X supported by $[0, 1]$ and that $S_N := \{X_1, \cdots X_N\}$ is a N-sample of X, giving rise to the e.d.f. $F_N(x)$. After Vitale (1975), who considered Bernstein density estimators for the first time, Babu et al. (2002) proposed an estimator

$\widetilde{F}_{N,m}$ of F, consisting in smoothing the random step function F_N:

$$\widetilde{F}_{N,m}(x) := \sum_{k=0}^{m} F_N(\frac{k}{m})\, w_{m,k}(x) = B_m[F_N]. \tag{7}$$

It is noteworthy that this estimator also smoothes another step function $F_{N,m}$ obtained by sub-sampling F_N, whose jump set is:

$$\left\{ (0,0)\,, \left(\frac{1}{m}, \frac{1}{N} \sum_{i=1}^{N} I\left(X_i < \frac{1}{m} \right) \right), \cdots, \left(\frac{k}{m}, \frac{1}{N} \sum_{i=1}^{N} I\left(X_i < \frac{k}{m} \right) \right), \cdots, (1,1) \right\}. \tag{8}$$

In fact the expression "$B_m[F_N]$" is slightly improper (F_N is not continuous) and should be replaced by "$B_m[\varphi_N]$", where φ_N should be some continuous function (piecewise linear, spline, *etc.*) interpolating the jump set (8), or should be obtained from a well-suited histogram (see for instance Birgé and Rozenholc (2002); Davies et al. (2009); Lugosi and Nobel (1996), and also Sections 6.1 & 6.2) by numerical integration. For the sake of simplicity, we will drop this refinement of no practical importance, except in the following proposition.

Proposition 5. *If F is a continuous d.f., the Bernstein operator is a contraction. More precisely,*

$$\|B_m\,[F]\| \leq \left(1 - \frac{1}{2^{m-1}} \right) \|F\|\,.$$

Consequently, if φ_N is a continuous estimate of F derived from F_N and such that $\varphi_N(0) = 0$ and $\varphi_N(1) = 1$, we can write:

$$\|B_m\,[\varphi_N - F]\| \leq \left(1 - \frac{1}{2^{m-1}} \right) \|\varphi_N - F\|\,.$$

Proof: see the appendix.

As a corollary, $B_m[F_N]$ inherits all the good asymptotic properties of the e.d.f. in the Chebyshev norm *(i.e.* in the Kolmogorov-Smirnov metric) (Servien , 2009; Leblanc , 2009; Babu et al. , 2002) because for large enough samples, F_N can always be closely approached by continuous functions.

Now, what about $\left\|\mathfrak{I}^I_{m;K}[F_N] - F\right\|$? The situation is much more intricate than in the classical Bernstein operator case, since $\mathfrak{I}^I_{m;K}$ is not a positive operator. Let us denote $\Delta_N := F_N - F$, and consider the sampled values $\delta_{m;N} := \left\{\Delta_N(0), \Delta_N\left(\frac{1}{m}\right) \cdots, \Delta_N\left(\frac{m-1}{m}\right), \Delta_N(1)\right\}$. We will also denote $H_{[m;K]}(\delta_{m;N})$ as the coordinates of $\mathfrak{I}^1_{m;K}[\Delta_N]$ in the equispaced Lagrange basis.

Proposition 6. *We can write:*

$$\left\|\mathfrak{I}^I_{m;K}[F_N] - F\right\| \leq \left\|H^I_{[m;K]}\right\|_\infty \|\delta_{m;N}\|_\infty \|\mathcal{L}_m\| + \left\|\mathfrak{I}^I_{m;K}[F] - F\right\|.$$

In addition:

$$\lim_{I\to\infty} \left\|\mathfrak{I}^I_{m;K}[F] - F\right\| \leq (1 + \|\mathcal{L}_m\|) \inf_{P\in\mathfrak{P}_n} \|P - F\|,$$

where $\|\mathcal{L}_m\| \approx \frac{2^m}{e\,m\,log(m)}$ *denotes the Lebesgue constant (Mills and Smith , 1992).*

Proof: see the appendix.

Remark 1. *To compute* $H^I_{[m;K]}$ *(except if* $I = K$*), we need to compute* $L\Pi^{-1}_{[m]}$ *(see the proof of Proposition 6). Since* $L\Pi_{[m]}$ *is ill-conditioned (like any change of polynomial basis),* $\left\|H^I_{[m;K]}\right\|_\infty$ *increases with* m*, just like* $\|\mathcal{L}_m\|$*, while* $\|\delta_{m;N}\|_\infty$ *clearly depends on the structure of* F*, and can be optimized*

(see Sections 6.1 & 6.2). On the other hand, a small value of m controls the possibly explosive behaviour of the undesirable (and essentially unknown) term $\left\|\mathfrak{I}_{m;K}^{I}[F] - F\right\|$ when I is big. Thus, Proposition 6 shows that the choice of (m, I) must result from a delicate tuning of these parameters.

The choice of the number of bins $m < N$ in formula 7 was previously discussed in Babu et al. (2002); Leblanc (2010, 2012a). Babu et al. (2002) proved the almost sure convergence of (7) when F is continuous, and gave conditions under which its rate of stochastic convergence can be determined, as well as the rate of convergence of the associated density estimator $\widetilde{f}_{N,m}(x) := \frac{d}{dx}\widetilde{F}_{N,m}(x)$ when F is differentiable with derivative $f := \frac{dF}{dx}$. More precisely, they proved that $\widetilde{f}_{N,m}$ almost surely converges towards f, under the condition $m = o(N/\ln(N))$. Furthermore, they inferred from simulations that the upper value $m = N/\ln(N)$ is indeed acceptable. But notice that $N \geq 100 \Rightarrow m > 21$. Thus, the numerical issues brought up is Section 5.1 will arise even with moderate sample size. Consequently, it is necessary to determine a number of bins $m \leq 21$ such that the associated partition of $[0, 1]$ is well-suited for F.

A similar problem was tackled by Manté (2012) but, instead of an e.d.f., the data consisted of a discretized distribution function $\{F(x_j), 0 \leq j \leq N\}$ sampled on an imposed mesh $0 < x_0 \leq x_1 < \ldots, < x_{N-1} \leq x_N < 1$. The method proposed by Manté (2012) consisted firstly in determining a submesh of size $n \leq N$ well-suited for Bernstein approximation and, secondly, in optimizing the number of iterations in formula 5, under the constraint that the associated density approximation $\widehat{f}_n^{(I^*)}$ is *bona fide* according to Gajek (1986), *i.e.* belongs to both the closed convex cone of positive functions $\mathcal{F}^+$

and the closed convex set $\mathcal{F}^1$ of functions integrating to one. A number of discretized distribution functions (*e.g.* grain size curves) were processed in that way (Manté , 2012; Manté and Stora , 2012). Sometimes we found $I^* = 1$; in such cases, the usual approximation cannot be improved by using Sevy's iteration scheme, because $\widehat{f}_n^{\,(2)} \notin \mathcal{F}^+ \cap \mathcal{F}^1$, while $\widehat{f}_n^{\,(1)} \in \mathcal{F}^+ \cap \mathcal{F}^1$. But we can indeed get finer trajectories by slowing down Sevy's acceleration method! We just have to supersede integers by rational numbers in formula (5), that is to say to use a K-fractional sequence (see Definition 2), whose resolution increases with K. This will be done in the next section.

But for the moment, the objective is to determine what number of bins is best-suited for a given data set. Since the upper value $m = N/\ln(N)$ is often too big to use in practice, we propose to lower it according to the structure of $F_N(x)$. Since $1/m$ can be considered as a kind of bandwidth (Leblanc , 2010), lowering m could cause oversmoothing, but one can expect that fractional iterations will offset this phenomenon. So, let us start with $m_0 := N/\ln(N)$, and consider the sequence of uniform meshes $\{U_m : 1 \leq m \leq m_0\}$ such that $U_m := \left\{\frac{i}{m},\, 0 \leq i \leq m\right\}$. We propose here a method to select $m^* \leq \min[m_0, 21]$ such that U_{m^*} is well-suited for F_N. By "well-suited", we mean that the same m^* should be obtained with a high probability from different samples of size N of X (stability), and that the step functions F_{N,m^*} and F_N should be close to each other (fidelity).

6.1. A stability/fidelity test

We first propose a criterion based on half-sampling (Stephens , 1978) and on a classical two-sample test. Suppose $N = 2M$ (if N is odd, get rid of an observation). From S_N, we randomly draw (without replacement) a

sample of size M of X, the learning sample S_M^L, and obtain the test sample $S_M^T := S_N \ominus S_M^L$. The subsamples S_M^L and S_M^T are independent, and the associated e.d.f.s will be denoted $F_M^L(x)$ and $F_M^T(x)$.

Even if the hypothesis $(\mathbf{H_0}) := F^L = F^T$ is actually true, due to sampling fluctuations (or to a descendant of Maxwell's demon), the e.d.f. of the subsamples S_M^L and S_M^T can be quite different, especially in the case of small samples. For instance, consider a N-sample of the uniform distribution: the probability of drawing a learning M-sample of numbers lower than 0.5 and a test M-sample of numbers greater than 0.5 is not null (with $M = 5$, it is 0.0625 and with $M = 10$, it is about 0.0020). This indeed depends upon the power of the test, and Stephens (1978) observed that the power of the half-sample goodness-of-fit test is uneven.

Consequently, we suggest to discard ill-suited subsamples such that the Kolmogorov-Smirnov random distance $D_{KS}\left(F_M^L, F_M^T\right) := \sup_{x\in[0,1]} \left|F_M^L(x) - F_M^T(x)\right|$ is excessive: in such a case, finding from the learning sample a mesh well-suited for the test sample is hopeless! Consider two samples of same size M of the same distribution, and the distribution-free statistics $D_{KS}\left(F^1, F^2\right)$ associated with the Kolmogorov-Smirnov homogeneity test. Gnedenko and Korolyuk (1951) obtained the exact distribution of this statistics; this probability measure $\mathcal{D}_M$ is defined by:

$P\left(\mathcal{D}_M \geq x\right) = 1$ when $x \leq 1/M$, $P\left(\mathcal{D}_M \geq x\right) = 0$ when $x \geq 1$ and

$$P\left(\mathcal{D}_M \geq x\right) = 1 - \sum_{i=-\lfloor 1/x \rfloor}^{\lfloor 1/x \rfloor} (-1)^i \frac{\begin{pmatrix} 2M \\ M - i\ \lfloor Mx \rfloor \end{pmatrix}}{\begin{pmatrix} 2M \\ M \end{pmatrix}},$$

when $x \in]1/M, 1[$ (see Der Megreditchian (1986) or Gnedenko and Korolyuk (1951)). In our case, because $(\mathbf{H_0})$ is true, we can write: $D_{KS}\left(F_M^L, F_M^T\right) \stackrel{\mathfrak{L}}{=} \mathcal{D}_M$.

Suppose now we randomly draw a pair of subsamples S_M^L and S_M^T from the data, and let $d := D_{KS}\left(F_M^L, F_M^T\right)$ be the computed distance between the associated e.d.f.s. If the p-value $P\left(\mathcal{D}_M \geq d\right)$ is big enough (≥ 0.95, say) the pair (L, T) is "good" since $(\mathbf{H_0})$ may be accepted with little risk. In this case we will use S_M^L to build a sequence $\left\{F_{M,m}^L : 1 \leq m \leq m_0\right\}$ of "coarsened" e.d.f.s, each $F_{M,m}^L$ being described by its jump set given by (8). If this isn't the case (*i.e.* if the pair (L, T) is "bad"), we draw another pair of subsamples, until $(\mathbf{H_0})$ is acceptable.

Suppose now that $(\mathbf{H_0})$ is accepted. It is noteworthy that the m^{th} coarsening process introduced above actually consists in replacing each $X_i^L \in [\frac{k}{m}, \frac{k+1}{m}[$ by the value $\frac{k}{m}$. In other words, this is a nonlinear transformation $\mathfrak{C}^m : [0, 1] \to U_m$ such that $x \in [\frac{k}{m}, \frac{k+1}{m}[\Rightarrow \mathfrak{C}^m(x) = \frac{k}{m}$. Consequently, we consider that the e.d.f. $F_{M,m}^L$ has been obtained from a sample of size M of the induced probability distribution, ${\mathfrak{C}^m}_* X$. Thus, computing $D_{KS}\left(F_{M,m}^L, F_M^L\right)$ should enable us to decide whether or not the hypothesis $(\mathbf{H_m}) : {\mathfrak{C}^m}_* X \stackrel{\mathfrak{L}}{=} X$ is acceptable, *i.e.* whether or not the coarsening significantly alters the data. But, since both these e.d.f.s are based on the same learning sample, testing this hypothesis from $D_{KS}\left(F_{M,m}^L, F_M^L\right)$ is impossible. On the other hand, using $D_{KS}\left(F_{M,m}^L, F_M^T\right)$ is straightforward, since S_M^L and S_M^T are independent: to accept or reject $(\mathbf{H_m})$, we just have to test whether or not the computed distance $D_{KS}\left(F_{M,m}^L, F_M^T\right)$ is an unlikely observation of $\mathcal{D}_M$.

We could compute all the distances $\left\{D_{KS}\left(F_{M,m}^L, F_M^T\right), 1 \leq m \leq m_0\right\}$ from

some good pair (L, T) and scan the corresponding list of p-values; acceptable values of m will be those for which the p-value is big enough (≥ 0.90, say). But since such random lists are highly fluctuating, it seems preferable to perform a reasonable number *(e.g.* 50) of good trials (such that $(\mathbf{H_0})$ is acceptable), and to summarize the obtained 50 lists of p-values by the associated m_0 box-plots (see the upper panel of Figures 2 & 5).

The reader can see in the upper panel of Figure 2, for instance, that p-values corresponding to $m < 7$ are very small. Consequently, such a coarsening would deeply alter the histogram structure. On the contrary, for $m > 15$ most p-values are greater than 0.5 and we can conclude that such a coarsening is quite acceptable.

6.2. A complementary fidelity criterion

We just proposed a method for obtaining a list of acceptable numbers of bins in histograms, but there are generally several candidates. To select one of them, we proceed with the complete sample S_N. This time, we compute the list of Hausdorff distances $\{d_{\mathcal{H}}(F_{N,m}, F_N), 1 \leq m \leq m_0\}$, which quantify the similarity of successive coarsened distributions with the complete e.d.f. (see the lower panel of Figures 2 & 5). Notice that these coarsened distributions are tightly associated with the classical estimator (Vitale , 1975; Babu et al. , 2002; Bouezmarni and Rolin , 2007; Leblanc , 2010, 2012a,b) through Eq. (7).

Remark 2. *The choice of the Hausdorff distance is supported by the works of Beer (1982) and Cuevas and Fraiman (1998) : $d_{\mathcal{H}}$ is a metric in the space of Upper Semi Continuous (USC) functions, and any d.f. is USC. Furthermore,*

if the theoretical d.f. F is continuous, the propositions $\|F_K - F\| \underset{K\to\infty}{\longrightarrow} 0$ *and* $d_{\mathcal{H}}(F_K, F) \underset{K\to\infty}{\longrightarrow} 0$ *are equivalent (Beer , 1982).*

To sum up, we will select $m^* \leq 21$ such that both $(\mathbf{H}_{\mathbf{m}^*})$ is acceptable and $d_{\mathcal{H}}(F_{N,m^*}, F_N)$ is visually small (see the lower panels of Figures 2 & 5).

7. Numerical illustrations

The method is tested hereunder on two data sets which can be found in the classical book of Silverman (1986).

7.1. The suicide Data

This data set is a classical benchmark in Density Estimation (Leblanc , 2010, 2012a; Silverman , 1986; Eilers and Marx , 1996), which consists of the duration (in days) of psychiatric treatment for 86 patients used in a study of suicide risks. These durations range between 1 and 737; consequently they must be rescaled to the unit interval with a transformation $\psi_{a,b}(x) := \frac{x-a}{b-a}$. Following Leblanc (2010, 2012a), we chose $a = 0$ and $b = 800$.

Notice that the integer closest to $86/\ln(86)$ is $m_0 = 19$, which was also the data-driven optimal choice found by Leblanc (2012a). Plots of the criteria proposed in the previous section are shown in Figure 2. On the upper panel are displayed the box-plots obtained with 50 good trials. The reader can see that there is generally not a very significant difference between $F^L_{43,m}$ and F^T_{43} for $m > 7$. Nevertheless, the lower panel of this figure shows that $d_{\mathcal{H}}(F_{43,m}, F_{43})$ is only small for $m \geq 13$, and that $m = 17$ gives an excellent approximation. In the end we chose $m^* = 18$, because all of the p-values corresponding to $\left\{D_{KS}\left(F^{L_i}_{43,18}, F^{T_i}_{43}\right) : 1 \leq i \leq 50\right\}$ were greater than 0.45,

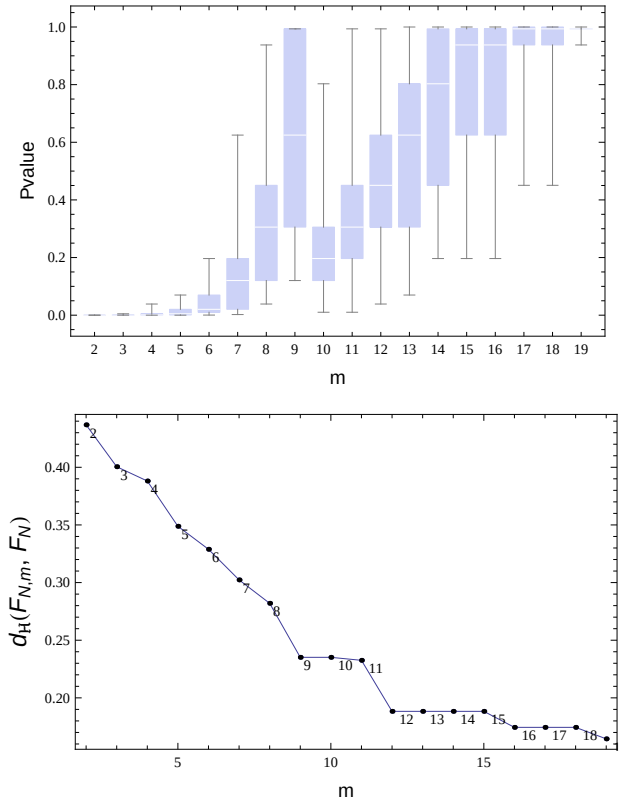

Figure 2:

while 25% of the p-values corresponding to $\left\{D_{KS}\left(F_{43,17}^{L_i}, F_{43}^{T_i}\right) : 1 \leq i \leq 50\right\}$ were lower than 0.80. The reader can see that this is also a satisfactory value from the fidelity point of view (lower panel of Figure 2).

Next, putting together the methodology of Manté (2012) and Definition 2, we first fix K, which determines the resolution of the discrete trajectory:

$$\left\{\mathfrak{P}_{m^*-1} \ni \widehat{f_{m^*}}^{(K+j)} := \frac{d\mathfrak{I}_{m^*;K}^{K+j}[F_{N,m^*}](x)}{dx},\ 0 \leq j\right\}$$

associated with the K^{th} "root" of the restricted operator $\overline{G_{m^*}}^{(1/K)}$. This trajectory consists in a sequence of polynomials, computed through Proposition 4. Remember that $\forall\,(m,K)\,,\,\mathfrak{I}_{m;K}^{K}\left[f\right] = \mathfrak{I}_{m}^{1}\left(f\right) = B_m\left[f\right]$. Consequently,

such a trajectory which starts in $\mathcal{F}^+ \cap \mathcal{F}^1$ ($\widehat{f_m}^{(K)} = \frac{d\widetilde{F}_{N,m}(x)}{dx} = \widetilde{f}_{N,m}(x)$ is always *bona fide*) and progressively get out of this closed convex set (in general, $\frac{d\mathfrak{I}^{\infty}_{m;K}[F_{N,m^*}](x)}{dx} = \frac{d\mathcal{L}_m\left[F_{N,m^*}\right](x)}{dx} \notin \mathcal{F}^+ \cap \mathcal{F}^1$). Thus, once m^* has been determined, it is quite natural to search for the first $I^* > K$ such that $\widehat{f_{m^*}}^{(I^*)}$ belongs to $\mathcal{F}^+ \cap \mathcal{F}^1$ while $\widehat{f_{m^*}}^{(I^*+1)}$ doesn't. For that purpose, we proposed (Manté , 2012) to control the graph of $\widehat{f_{m^*}}^{(i)}$ through two "stresses" : the positivity stress

$$\pi(i) := \int_0^1 \left(\left| \widehat{f_{m^*}}^{(i)} \right| - \widehat{f_{m^*}}^{(i)} \right)(x)dx, \tag{9}$$

and the unit total mass stress

$$\nu(i) := \int_0^1 \left(\widehat{f_{m^*}}^{(i)} + \left| \widehat{f_{m^*}}^{(i)} \right| \right)(x)dx - 2. \tag{10}$$

The approximation $\widehat{f_{m^*}}^{(i)}$ is *bona fide* if and only if both of these stresses are null.

Remark 3. *As in (Manté , 2012), all of the computations are made in the Bernstein basis.*

We fixed $K = 10$, and plotted stresses (9) and (10) in Figure 3, together with the Kolmogorov distance (in percents)

$$K.D.(i) := 100 \sup_{x\in[0,1]} \left| \int_0^x \widehat{f_{m^*}}^{(K+i)}(t)dt - F_N(x) \right|.$$

Since in our case $\int_0^1 \left| \widehat{f_{m^*}}^{(K+i)} \right|(x)dx \approx 1$, the curves π and ν are indeed very similar to each other. One can see that for this data set, $\widehat{f_{m^*}}^{(K+i)} \not\geqslant 0$, except

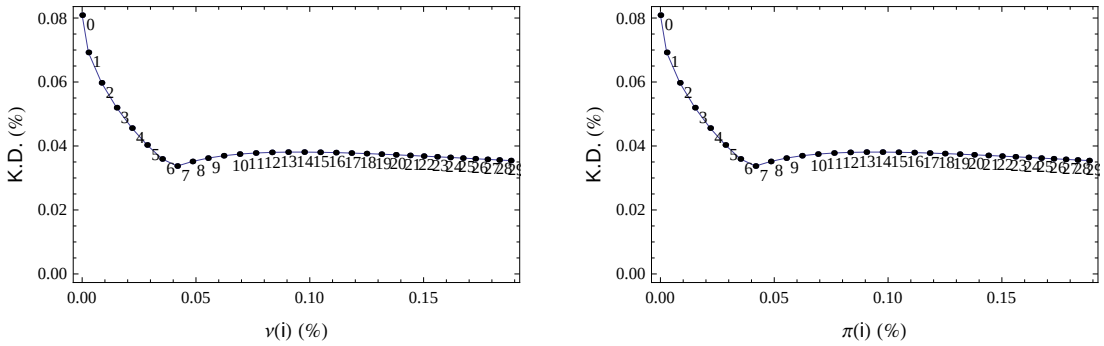

Figure 3:

for $i = 0$. The best fit is attained at the 7^{th} iteration, corresponding to the fractional power $r^* = \frac{10+7}{10} = 1.7$. Notice that the values of the stresses (around 0.0004) are very small; thus, the estimated density is practically *bona fide*. The obtained estimates are finally displayed in Figure 4. On the upper panel of the figure, we plotted :

1. the e.d.f. and its Gnedenko confidence bands with coverage probability 0.95 (red) and 0.999 (green)
2. the Bernstein estimators : $B_{m_0}[F_{N,m_0}]$ of Babu et al. (2002) (of degree 19) and $B_{m^*}[F_{N,m^*}]$ (of degree 18); the reader can see that they are very close to each other
3. the proposed estimator, $\mathfrak{I}^{K+I^*}_{m^*;K}[F_{N,m^*}]$, which is also a polynomial of degree 18.

On the lower panel, we plotted the three corresponding density estimators. Please note that the exponential aspect of these three densities have been highlighted in previous studies (Eilers and Marx , 1996; Leblanc , 2010, 2012a). Nevertheless, kernel (Silverman , 1986) or spline estimators (Eilers and Marx (1996), p. 99) behaved differently from ours near zero. Such differ-

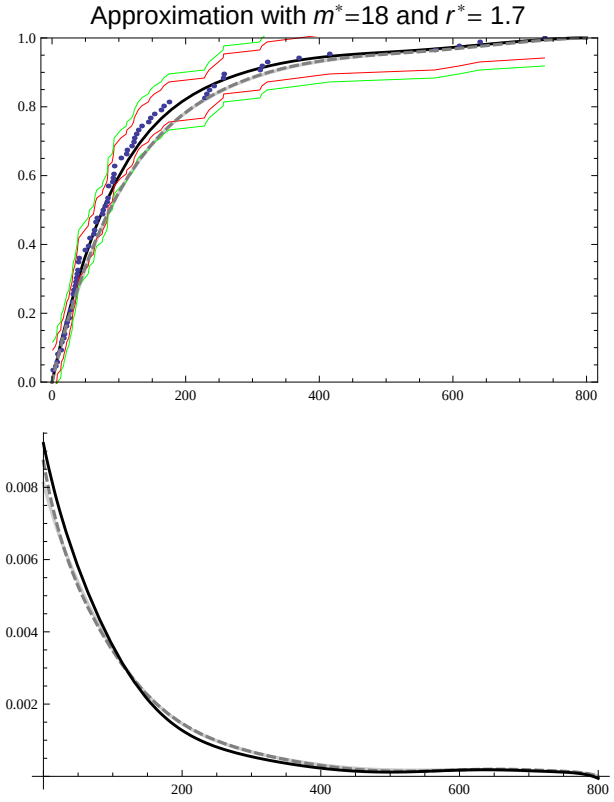

Figure 4:

ences likely come from the fact that, contrary to most other methods, Bernstein estimators are well-behaved near boundaries (Bouezmarni and Rolin , 2007; Leblanc , 2012b).

7.2. The Old faithful data

This data set consists of 107 eruption lengths of the Old Faithful geyser, situated in the Yellowstone National Park. These lengths range between 1.67 and 4.93 minutes. Thus, we embedded these data in the interval $[1.5, 5]$ and rescaled them to the unit interval (following Leblanc (2010)). In this case, $m_0 = 23$ although Figure 5 shows that choosing $m^* = 18$ is quite reasonable:

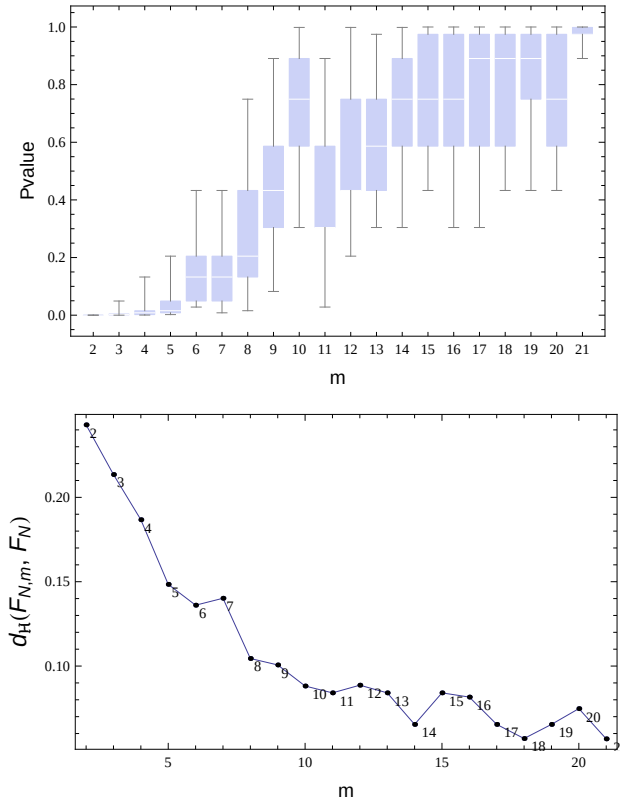

Figure 5:

from the stability point of view (see the upper panel) greater values are not better, except of course for the choice $m^* = 23$ which is too big. In addition, choosing $m^* = 18$ satisfies the fidelity criterion (see the lower panel of Figure 5).

With $K = 10$ we found $I^* = 9$, and obtained the d.f. and density estimations plotted in Figure 6. Notice that in this case, the Babu et al. (2002) density estimator of degree 22 is close to the derivative of $B_{m^*}[F_{N,m^*}]$ (of degree 18) while our estimate is rather different: it is similar to estimates obtained by various authors with kernel (Silverman (1986) p.17, S.T. Chiu (1991) p. 1897 and Sain and Scott (1996)) or spline estimators (Eilers and

Marx (1996), pp. 99 and 118). The Bernstein estimator of Leblanc (2010) was more "bumpy", probably because it belonged to $\mathfrak{P}_{65}$, while ours lies in $\mathfrak{P}_{18}$. Nevertheless it is worthwhile to take the comparison of these results a bit further.

Following the pioneering work of Vitale (1975), Leblanc (2010) proved that $\widetilde{f}_{N,m}(x) := \frac{d}{dx}\widetilde{F}_{N,m}(x)$ is biased, and proposed instead the biased-corrected estimator $2\,\widetilde{f}_{N,m}(x) - \widetilde{f}_{N,m/2}(x)$. He was inspired by a paper by Politis and Romano (1995) which was dedicated to spectral density estimation. Roughly speaking, the method of Politis & Romano consisted in reducing the bias of the Bartlett spectral estimator $\hat{f}\,(\omega)$ by computing instead $2\,\hat{f}\,(\omega) - \bar{f}\,(\omega)$, where $\bar{f}\,(\omega)$ was an over-smoothed Bartlett spectral estimator. Now, in the setting of density estimation and with the notations of Leblanc (2010), if m has been well-chosen, $\widetilde{f}_{N,m}$ is a pertinent estimator while $\widetilde{f}_{N,m/2}$ is necessarily an oversmoothed estimator. Consider now our estimate $\widehat{f_{m^*}}^{(2)} = \frac{d\mathfrak{I}^2_{m^*}[F_{N,m^*}](x)}{dx}$, and notice that (see formula 5):

$$\mathfrak{I}^2_{m^*}[F_{N,m^*}] = \left(1 - (1 - B_{m^*})^2\right)[F_{N,m^*}] = \left(2\,B_{m^*} - B^2_{m^*}\right)[F_{N,m^*}].$$

It is well-known (Cooper and Waldron , 2000) that, because of the eigenvalues of B_n, iterated operators B_n^k act as filters, such that $\lim_{k\to\infty} B_n^k\,[f] = \mathcal{L}_1\,[f]$. Thus, $\widehat{f_{m^*}}^{(2)}$ has the same structure as the biased-corrected estimator of Leblanc (2010), except that the over-smoothed component $B^2_{m^*}[F_{N,m^*}]$ is built differently. This might explain why for a lot of data sets studied the optimal fractional iteration number $r^* := \frac{K+I^*}{K}$ was close to 2.

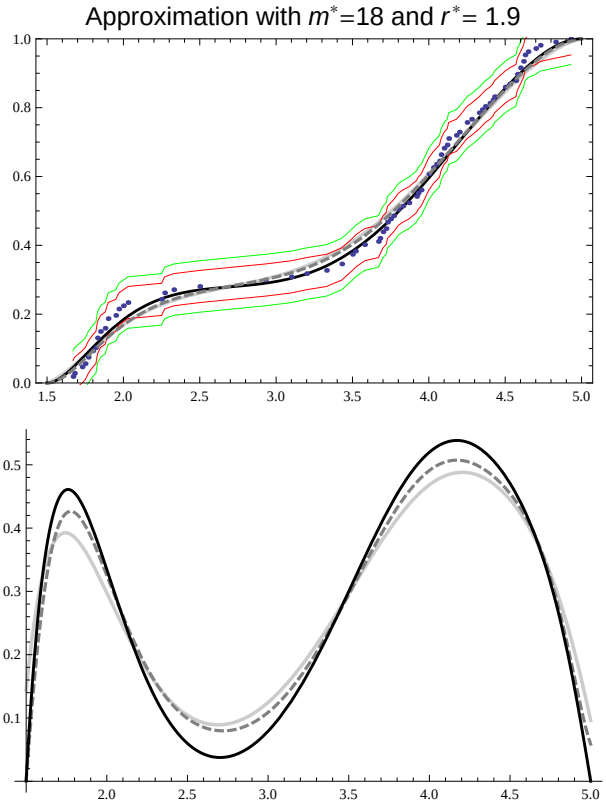
Approximation with m^*=18 and r^*= 1.9
1.0
0.8
0.6
0.4
0.2
0.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
0.5
0.4
0.3
0.2
0.1
2.0
2.5
3.0
3.5
4.0
4.5
5.0

Figure 6:

8. Discussion

In this paper we propose an original method for estimating distribution functions and densities with Bernstein polynomials. On the one hand, we take advantage of results about the eigenstructure of the Bernstein operator to improve Sevy's convergence acceleration method. On the other hand, we work out an original adaptative method for choosing the number of bins m of a regular histogram. As Birgé and Rozenholc (2002) noticed: this is an old and still open problem. **In the setting of Bernstein estimation** of distribution functions and densities, Babu et al. (2002) proposed the upper value $m_0 := N/\ln(N)$ as an "acceptable" solution to this problem, even if one should theoretically choose $m = o(N/\ln(N))$. In theory, the number of bins should not be the same for fitting both the d.f. and the density. Leblanc (2010, 2012a) proved that (asymptotically) $m^* = O\left(N^{2/5}\right)$ when one focusses on density estimation, and $m^* = O\left(N^{2/3}\right)$ when one focusses on d.f. estimation. As for Babu et al. (2002), they recommended choosing $m^* = o\left(\frac{N}{\log(N)}\right)$ in the former case and $m^* = O\left(\left(\frac{N}{\log(N)}\right)^2\right)$ in the latter one. Thus the density estimator should be built from a smoother d.f. estimator than the optimal d.f. estimator itself. A similar result was proved by Hjort and Walker (2001) regarding kernel density estimation. This is probably due to the fact that, roughly speaking, the differentiation operator is a high-pass filter whose action must generally be balanced by smoothing.

Our two-step method takes both functions into account: $m^* \leq m_0$ - well-suited for density estimation (Babu et al. , 2002) is first tuned according to the structure of the e.d.f., and then r^* is tuned according to the density which should be *bona fide*. These steps cannot be interchanged because m^*

determines the best subspace while r^* corresponds to the optimal number of iterations of an operator acting **inside** this subspace. Thus, simultaneous bivariate optimization is unnatural.

It is noteworthy that m_0 was independently proposed by Birgé and Rozenholc (2002) as an upper number of classes **in the setting of automatic histograms construction**. Unfortunately, m_0 is generally too big for us (numerical issues), but we stress that big values of m are indeed linked to the sluggish convergence of Bernstein approximations. For instance, the Voronovsky theorem (Davis , 1963) proves that the rate of uniform convergence of the Bernstein approximation of a twice differentiable function is $O(1/m)$, while the rate of convergence of the best polynomial approximation is much better: it is $\varepsilon(m)\ln(m)/m^2$, with $\lim_{m\to\infty}\varepsilon(m)=0$ (Laurent (1972), p. 303). In the special case of distribution functions, we proved (Manté , 2012) that one can expect only $O(1/\sqrt{m})$ as a rate of convergence of the Bernstein approximation. However, it is possible to compensate for this drawback thanks to the acceleration of convergence method used here: similar estimates can be obtained either with a small number of iterations in a large space of polynomials or else with a large number of iterations in a smaller space (compare Figures 6 & 7; see also Remark 1).

Contrary to the method of Babu et al. (2002), the method of Birgé and Rozenholc (2002) gives an optimal value m_{BR} which is generally too small for our purpose. For instance, in the case of the suicide data, it gives the single optimum $m_{BR}=6\ll m_0=19$. In the case of the geyser data, one should choose $m_{BR}=9\ll m_0=23$ (this is the greater optimum of this criterion). Thus, the associated density estimators are of very low degree (respectively

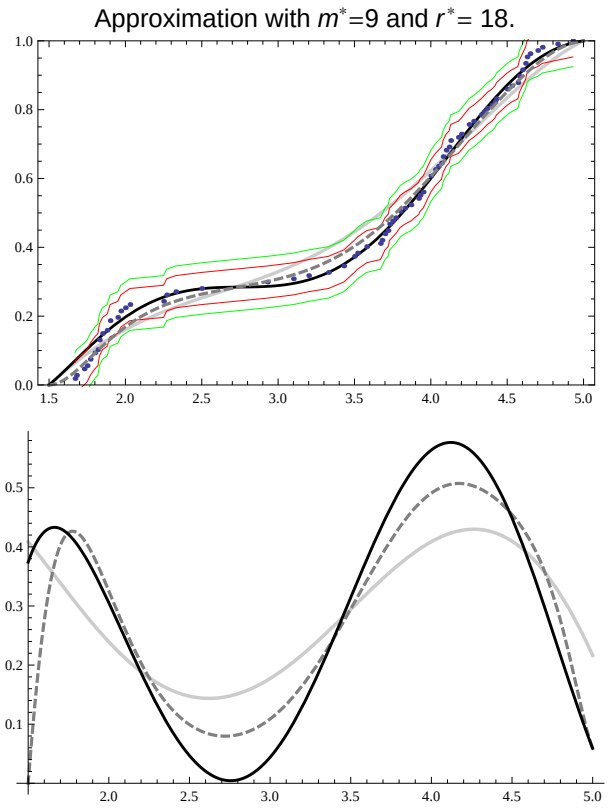
Approximation with m^*=9 and r^*= 18.
1.0
0.8
0.6
0.4
0.2
0.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
0.5
0.4
0.3
0.2
0.1
2.0
2.5
3.0
3.5
4.0
4.5
5.0

Figure 7:

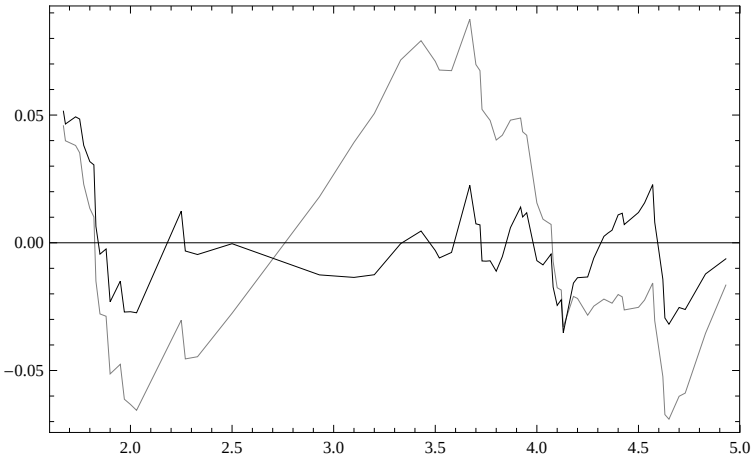

Figure 8:

5 and 8). This can be offset by using a large iteration number: indeed, in the first case we could find a good approximation with $r^* = 5$ while in the second case $r^* = 18$ would be correct (see Figure 7). On the contrary, the classical Bernstein estimators, of the same degree, (gray curves on Figure 7) are extremely oversmoothed. This is displayed in more detail on Figure 8, where we plotted the difference between the raw e.d.f. F_{107} and its Bernstein estimate $B_9\left[F_{107,9}\right]$, superimposed with $F_{107} - \mathfrak{I}_{9;1}^{20}[F_{107,9}]$, also of degree 9. Thus, using K-fractional Sevy approximation sequence, it is possible to obtain satisfactory estimators, even with low-degree polynomials, when the density is smooth enough! Clearly, in the case of the Old Faithful data, degree 9 is not enough for regions of strong curvature (both the extremities of the curves), but it is enough for weakly curved regions (see Figure 8) when the number of iterations is sufficient.

The main issue is indeed the condition that $m \leq 21$, which stems from the numerical problems raised in Section 5.1. These difficulties are due to the ill-conditioning of matrices involved in the expression: $Mat\left(\overset{\circ}{B}_n; L_n, W_n\right) =$

$\Pi W_{[n]} \circ \Lambda_{[n]} \circ L\Pi_{[n]}$. At first sight, one cannot do anything against the curse of ill-conditioning... Nevertheless, we stressed in Section 5.1 the excessive computational cost of the polynomials calculated from the complicated recurrence formula of Cooper and Waldron (2000). The observed loss of accuracy is probably due to the complexity of these calculations.

Please note that on the one hand $Mat\left(\overset{\circ}{B}_n; W_n, W_n\right) = LW_{[n]}$ and $Mat\left(\overset{\circ}{B}_n; L_n, L_n\right) = LW_{[n]}$ (see Lemma 2) while on the other hand, $B_n[f] = \sum_{j=0}^{n} \lambda_j^{[n]} \pi_j^{[n]} \otimes \pi_j^{*[n]} \left(\mathcal{L}_n[f]\right)$. As a consequence, $\Pi_{[n]}$ can be calculated by polynomial interpolation of the eigenvectors of $LW_{[n]}$. This is quicker and much more numerically stable, and makes it possible to go beyond the limit $m = 21$ (for further details, see (Manté , 2014)).

In spite of this exciting perspective, one should be warned that, no matter what, m is necessarily bounded by $\max[m_0, M]$ (where M is to be determined) because

- for large values of this upper value, the method proposed in Section 6.1 for determining m^* would be excessively computationally expensive
- m^* itself must be bounded, because of the ill-conditioning problems, and more generally because "it is impractical to employ polynomials with degrees running to hundreds or thousands in "real-world" problems" (Farouki , 2012).

Indeed, the proposed method is heuristic, constructive, and well-suited for moderate sample sizes in its current state.

9. Figures captions

Figure 1: plot of the logarithm of the norms $\left\{\left\|\Pi W_{[n]} \circ \Lambda_{[n]} \circ L\Pi_{[n]} - I_n\right\|_\infty, 1 \leq n \leq 35\right\}$, as an indicator of loss of numerical accuracy in the computation of $Mat\left(\overset{\circ}{B}_n; L_n, W_n\right)$, due to transformation matrices.

Figure 2: (suicide data). Upper panel: box-plots of the p-values of

$$\left\{D_{KS}\left(F_{43,m}^{L_i}, F_{43}^{T_i}\right),\ 1 \leq i \leq 50;\ 1 \leq m \leq m_0 = 19\right\},$$

assuming that each $D_{KS}\left(F_{43,m}^{L}, F_{43}^{T}\right)$ obeys $\mathcal{D}_{43}$. Lower panel: plot of the Hausdorff distances $\left\{d_{\mathcal{H}}\left(F_{86,m}, F_{86}\right), 1 \leq m \leq 19\right\}$.

Figure 3: (suicide data) Plot, with the step 1/10, of the stresses $\nu(i)$ and $\pi(i)$ characterizing $\widehat{f_{18}}^{(i)}$, against the Kolmogorov distance (in percents) $K.D.(i)$.

Figure 4: (suicide data). Upper panel: plot of both the Bernstein estimators $B_{19}\left[F_{86,19}\right]$ of Babu et al. (2002) (dashed gray) and $B_{18}\left[F_{86,18}\right]$ (gray) superimposed to the proposed one, $\mathfrak{I}_{19;10}^{17}[F_{86,19}]$ (black), and to F_{86} (dots) and the associated Gnedenko confidence bands with coverage probability 0.95 (red) and 0.999 (green). Lower panel: the density estimates, with the same graphic directives as in the upper panel.

Figure 5: (geyser data). Upper panel: box-plots of the p-values of

$$\left\{D_{KS}\left(F_{53,m}^{L_i}, F_{53}^{T_i}\right),\ 1 \leq i \leq 50;\ 1 \leq m \leq m_0 = 23\right\},$$

assuming that each $D_{KS}\left(F_{53,m}^{L}, F_{53}^{T}\right)$obeys $\mathcal{D}_{53}$. Lower panel: plot of the Hausdorff distances $\left\{d_{\mathcal{H}}\left(F_{107,m}, F_{107}\right), 1 \leq m \leq 23\right\}$.

Figure 6: (geyser data). Upper panel: plot of both the Bernstein estimators $B_{23}[F_{107,23}]$ of Babu et al. (2002) (dashed gray) and $B_{18}[F_{107,18}]$ (gray) superimposed to the proposed one, $\mathfrak{I}^{19}_{18;10}[F_{107,18}]$ (black), and to F_{107} (dots) and the associated Gnedenko confidence bands with coverage probability 0.95 (red) and 0.999 (green). Lower panel: the density estimates, with the same graphic directives as in the upper panel.

Figure 7: (geyser data). Upper panel: plot of both the Bernstein estimators $B_{23}[F_{107,23}]$ of Babu et al. (2002) (dashed gray) and $B_{9}[F_{107,9}]$ (gray) superimposed to the proposed one, $\mathfrak{I}^{17}_{9;1}[F_{107,9}]$ (black), and to F_{107} (dots) and the associated Gnedenko confidence bands with coverage probability 0.95 (red) and 0.999 (green). Lower panel: the density estimates, with the same graphic directives as in the upper panel.

Figure 8: (geyser data). Plot of the differences $F_{107} - B_{9}[F_{107,9}]$ (gray) and $F_{107} - \mathfrak{I}^{20}_{9;1}[F_{107,9}]$ (black) between the raw d.f. and iterated Bernstein estimates.

Acknowledgements

The author is very grateful to the referees for their numerous and helpful comments and suggestions, and to Starrlight Augustine for greatly improving the English text.

Appendix: proofs of intermediate results

Proof of Corollary 1

Since $B_n\left[\left(f-\mathcal{L}_n\left[f\right]\right)\right]=0$, we can restrict ourselves to the case where $f\in\mathfrak{P}_n=\mathcal{R}\left(B_n\right)$. Since $\left\{\pi_0^{[n]},\cdots,\pi_n^{[n]}\right\}$ is a basis of this space, and $\left\{\mu_0^{[n]},\cdots,\mu_n^{[n]}\right\}$ is a basis of $\mathfrak{P}_n^*$ we can write, using the standard notation $u\otimes v^*\left(w\right):=u\left\langle v^*,w\right\rangle$ (Bowen and Wang , 1976):

$$B_n\left[f\right]=\overset{\circ}{B}_n\circ\mathcal{L}_n\left[f\right]=\sum_{j=0}^{n}\lambda_j^{[n]}\pi_j^{[n]}\left\langle\mu_j^{[n]},\mathcal{L}_n\left[f\right]\right\rangle=\sum_{j=0}^{n}\lambda_j^{[n]}\pi_j^{[n]}\otimes\pi_j^{*[n]}\left(\mathcal{L}_n\left[f\right]\right)$$

□

Proof of Lemma 3

Notice first that:

$$\left(1-\left(1-B_n\right)^I\right)=\sum_{k=1}^{I}\left(-1\right)^{k-1}\left(\begin{array}{c}I\\k\end{array}\right)B_n^k.$$

Then, thanks to Lemma 1, we can write:

$$\sum_{k=1}^{I}\left(-1\right)^{k-1}\left(\begin{array}{c}I\\k\end{array}\right)B_n^k=\left(\sum_{k=1}^{I}\left(-1\right)^{k-1}\left(\begin{array}{c}I\\k\end{array}\right)\overset{\circ}{B}_n^k\right)\circ\mathcal{L}_n=\left(1-\left(1-\overset{\circ}{B}_n\right)^I\right)\circ\mathcal{L}_n$$

□

Proof of Proposition 2

Remember that $\overset{\circ}{G}_n$ can be considered as a symmetrical bilinear application, whose matrix can be written $Mat\left(\overset{\circ}{G}_n;E_{[n]},E_{[n]}\right)=Q_{[n]}\,\Gamma_{[n]}\,Q_{[n]}^t$, where $Q_{[n]}$ is orthogonal and $\Gamma_{[n]}$ is the diagonal matrix associated with the vector $(0,0,1/n,(3n-2)/n^2,\cdots,1-n!/n^n)$. Eliminating its two first

eigenfunction (basis of $Ker(\overset{\circ}{G}_n)$) amounts to restrict this operator to $\overline{\mathfrak{P}_n}$. As a consequence, we may write $Mat\left(\overline{G_n}; \overline{E_{[n]}}, \overline{E_{[n]}}\right) = \overline{Q_{[n]}}\ \ \overline{\Gamma_{[n]}}\ \ \overline{Q_{[n]}}^t$, where $\overline{Q_{[n]}}$ is orthogonal, and $\overline{\Gamma_{[n]}}$ is the diagonal matrix associated with the vector $(1/n, (3n-2)/n^2, \cdots, 1-n!/n^n)$. Then, we have first: $\log\left(\overline{G_n}\right) = \sum_{k=1}^{\infty} \frac{\overline{B_n}^k}{k}$ and, because $Mat\left(\overline{B_n}^k; \overline{E_{[n]}}, \overline{E_{[n]}}\right) = \overline{Q_{[n]}}\ \ \overline{\Lambda_{[n]}}^k\ \ \overline{Q_{[n]}}^t$,

$$Mat\left(\log\left(\overline{G_n}\right); \overline{E_{[n]}}, \overline{E_{[n]}}\right) = \overline{Q_{[n]}}\ \sum_{k\geq 1} \frac{\overline{\Lambda_{[n]}}^k}{k}\ \overline{Q_{[n]}}^t = \overline{Q_{[n]}}\ \ \overline{\Delta_{[n]}}\ \ \overline{Q_{[n]}}^t,$$

where $\overline{\Delta_{[n]}}$ is the diagonal matrix associated with the vector

$$\left(\sum_{k\geq 1} \frac{(1-1/n)^k}{k}, \sum_{k\geq 1} \frac{\left(1-(3n-2)/n^2\right)^k}{k}, \cdots, \sum_{k\geq 1} \frac{(n!/n^n)^k}{k}\right) = \left(\log\left(\frac{1}{n}\right), \log\left(\frac{(3n-2)}{n^2}\right), \cdots, \log\left(1-\frac{n!}{n^n}\right)\right).$$

Proceeding the same way with the exponential operator, we find:

$$Mat\left(\exp\left(\alpha \log\left(\overline{G_n}\right)\right); \overline{E_{[n]}}, \overline{E_{[n]}}\right) = \overline{Q_{[n]}}\ \ \overline{\Gamma_{[n]}}^{(\alpha)}\ \ \overline{Q_{[n]}}^t$$

□

Proof of Proposition 3

We demonstrate this proposition by induction. Firstly, note that $\mathfrak{I}_n^1(P) = B_n(P) = P_1 + B_n\left(\bar{P}\right) = P_1 + \mathfrak{I}_n^1\left(\bar{P}\right)$; let us now suppose that for some $k > 1$, $\mathfrak{I}_n^k(P) = P_1 + \mathfrak{I}_n^k\left(\bar{P}\right)$. Then:

$$\begin{aligned}
\mathfrak{I}_n^{k+1}(P) &= P - (1-B_n)^{k+1}(P), \\
&= P_1 + \bar{P} - (1-B_n)^k\left((1-B_n)(P)\right), \\
&= P_1 + \bar{P} - (1-B_n)^k\left((1-B_n)\left(\bar{P}\right)\right), \\
&= P_1 + \mathfrak{I}_n^{k+1}\left(\bar{P}\right) \quad \square
\end{aligned}$$

Proof of Proposition 5

Let us first define the closed subset $C_{a,b} := \{f \in C[0,1] \mid f(0) = a,\ f(1) = b\}$. If F is a d.f. supported by the unit interval, $F \in C_{0,1}$ while, if F_N a e.d.f. associated with a N-sample, $(F_N - F) \in C_{0,0}$.

Both the inequalities claimed result from the fact that the restriction $B_m \mid_{C_{a,b}}: C_{a,b} \to C_{a,b}$ is a contraction. More precisely, Rus (2004) proved that for all $f, g \in C_{a,b}$, $\|B_m[f-g]\| \leq \left(1 - \frac{1}{2^{m-1}}\right)\|f-g\|$ $\square$

Proof of Proposition 6

Notice first that:

$$\left\|\mathfrak{I}^I_{m;K}[F_N] - F\right\| \leq \left\|\mathfrak{I}^I_{m;K}[\Delta_N]\right\| + \left\|\mathfrak{I}^I_{m;K}[F] - F\right\|,$$

and, because of Theorem 2, we have the asymptotical relation (Laurent , 1972, p. 303):

$$\lim_{I\to\infty}\left\|\mathfrak{I}^I_{m;K}[F] - F\right\| \leq (1 + \|\mathcal{L}_m\|)\inf_{P\in\mathfrak{P}_n}\|P - F\|.$$

From another side, consider the vector:

$$\delta_{m;N} := \left\{\Delta_N(0),\ \Delta_N\left(\frac{1}{m}\right)\cdots, \Delta_N\left(\frac{m-1}{m}\right),\ \Delta_N(1)\right\}.$$

The coordinates of $\mathfrak{I}^I_{m;K}[\Delta_N]$ in the Lagrange basis are given by $H^I_{[m;K]}(\delta_{m;N}) := L\Pi^{-1}_{[m]} \circ \Lambda^{(I/K)}_{[m]} \circ L\Pi_{[m]} \circ \delta_{m;N}$ (see Diagram (3)). In particular:

$$H^K_{[m;K]}(\delta_{m;N}) = L\Pi^{-1}_{[m]} \circ \Lambda^{(1)}_{[m]} \circ L\Pi_{[m]} \circ \delta_{m;N} = WL_{[m]} \circ \delta_{m;N} = B_m[\Delta_N],$$

and of course:

$$\lim_{I\to\infty} \mathfrak{I}^I_{m;K}\left[\Delta_N\right] = L\Pi^{-1}_{[m]} \circ \Lambda^{(\infty)}_{[m]} \circ L\Pi_{[m]} \circ \delta_{m;N} = \delta_{m;N} = \mathcal{L}_m\left(\Delta_N\right).$$

Thus, for each $I \geq K+1$, $\left\|\mathfrak{I}^I_{m;K}\left[\Delta_N\right]\right\|$ depends on the matrix $H^I_{[m;K]}$: $\delta_{m;N} \mapsto \delta^I_{m;N}$, such that $H^I_{[m;K]}\left(\delta_{m;N}\right) = \sum_{j=0}^{n} \left(\delta^I_{m;N}\right)_j \ell_{m,j}\left(x\right)$, and we have:

$$\left\|\mathfrak{I}^I_{m;K}\left[\Delta_N\right]\right\| \leq \left\|H^I_{[m;K]}\right\|_\infty \left\|\delta_{m;N}\right\|_\infty \left\|\mathcal{L}_m\right\|$$

□

References

G. J. Babu, A. J. Canty and Y. P. Chaubey. Application of Bernstein polynomials for smooth estimation of a distribution and density function. Journal of Statistical Planning and Inference, 105(2002), 377-392.

G. Beer. Upper semicontinuous functions and the Stone approximation theorem. Journal of Approximation Theory, 34(1982), 1-11.

S. N. Bernstein. Démonstration du théorème de Weierstrass fondée sur le calcul des probabilités. Commun. Soc. Math. Kharkov,13(1912), 1-2.

L. Birgé and Y. Rozenholc. How many bins should be put in a regular histogram? ESAIM: Probability and Statistics, 10(2002), 24-45.

T. Bouezmarni and J.M. Rolin. Bernstein estimator for unbounded density function. Journal of Nonparametric Statistics, 19, 3(2007), 145-161.

R. Bowen and C.C. Wang. Introduction to Vectors and Tensors Volume 1: Linear and Multilinear Algebra (Mathematical Concepts and Methods in Science and Engineering) , Springer, New York,1976.

S.T. Chiu. Bandwidth selection for kernel density estimation. The Annals of Statistics, 19, 4(1991), 1883-1905.

S. Cooper and S. Waldron. The eigenstructure of the Bernstein operator. Journal of Approximation Theory, 105(2000), 133-165.

A. Cuevas, R. Fraiman. On visual distances in density estimation : the Hausdorff choice. Statistics & Probability Letters, 40(1998), 333-341.

S. M. Curtis, S. K. Ghosh. A variable selection approach to monotonic regression with Bernstein polynomials. Journal of Applied Statistics, 38, 5(2011), 961-976.

L. Davies, U. Gather, D. Nordman and H. Weinert. A comparison of automatic histogram constructions. ESAIM: Probability and Statistics, 13 (2009), 181-196.

P.J. Davis. Interpolation and approximation, Blaisdell, New York, 1963.

C. de Boor. A practical guide to splines. Applied Mathematical Sciences, 27, Springer-Verlag, New York, 1978.

G. Der Megreditchian. Un test non paramétrique unilatéral de rupture d'homogénéité de "K" échantillons. Revue de Statistiques Appliquées, XXXIV, 1(1986), 45-60.

P. H. C. Eilers and B. D. Marx. Flexible smoothing with B-splines and penalties (with discussion). Statistical Science,11, 2(1996), 89-121.

R. T. Farouki. On the stability of transformations between power and Bernstein polynomials forms. Computer Aided Geometric Design, 8(1991), 29-36.

R.T. Farouki. The Bernstein polynomial basis: a centennial retrospective. Computer Aided Geometric Design, 29(2012), 379-419.

L. Gajek. On improving density estimators which are not bona fide functions. The Annals of Statistics, 14, 4(1986), 1612-1618.

B.V. Gnedenko and V.S. Korolyuk. On the maximum discrepancy between two empirical distribution functions. Dokl. Akad. Nauk. SSSR 80, 525-528; English translation: Select. Transl. Math. Statist. Probab. 1 (1961), 13-16.

T. Hermann. On the stability of polynomial transformations between Taylor, Bernstein and Hermite forms. Numerical Algorithms, 13(1996), 307-320.

N. L. Hjort and S. G. Walker. A note on kernel density estimators with optimal bandwidth. Statistics & Probability Letters, 54(2001), 153-159.

C. Impens, Stirling's series made easy. Amer. Math'l Monthly, 110(2003), 730-735.

T. Kato. Perturbation theory for linear operators. Springer-Verlag, Berlin, Heidelberg, 1995.

P.-J. Laurent. Approximation et optimisation. Enseignement des sciences, 13, Hermann, Paris, 1972.

A. Leblanc. Chung–Smirnov property for Bernstein estimators of distribution functions. Journal of Nonparametric Statistics, 21, 2(2009), 133-142.

A. Leblanc. A Bias-reduced approach to density estimation using Bernstein polynomials. Journal of Nonparametric Statistics, 22, 4(2010), 459–475.

A. Leblanc. On estimating distribution functions using Bernstein polynomials. Ann. Inst. Stat. Math., 64(2012), 919-943.

A. Leblanc. On the boundary properties of Bernstein polynomial estimators of density and distribution functions. Journal of Statistical Planning and Inference, 142 (2012), 2762-2778.

G. Lugosi and A. Nobel. Consistency of data-driven histogram methods for density estimation and classification. The Annals of Statistics, 24, 2(1996), 687-706.

C. Manté. Application of iterated Bernstein operators to distribution function and density approximation. Applied Mathematics and Computation, 218(2012), 9156-9168.

C. Manté and G. Stora. Functional PCA of measures for investigating the influence of bioturbation on sediment structure, in: Colubi, A., Fokianos, K., Kontoghiorghes, E. J., Gonzalez-Rodrigez, G. (Eds.), Proceedings of COMPSTAT 2012, pp. 531-542.

C. Manté. Density and Distribution Function estimation through iterates of fractional Bernstein Operators, in: Gilli, M., Gonzalez-Rodrigez, G., Nieto-Reyes, A. (Eds.), Proceedings of COMPSTAT 2014, pp. 335-342.

A. Marco and J. J. Martinez. Polynomial least squares fitting in the Bernstein basis. Linear Algebra and its Applications, 433(2010), 1254-1264.

T.M. Mills and S.J. Smith. The Lebesgue constant for Lagrange interpolation on equidistant nodes. Numerische Mathematik, 61(1992), 111-115.

S. Petrone. Random Bernstein Polynomials. Scandinavian Journal of Statistics, 26(1999), 373-393.

D.N. Politis and J.P. Romano. Bias-corrected nonparametric spectral estimation. J. Time Ser. Anal., 16 (1995), 67-103.

I. A. Rus. Iterates of Bernstein operators, via contraction principle. Journal of Mathematical Analysis and Applications, 292(2004), 259-261.

A. Sahai. An iterative algorithm for improved approximation by Bernstein's operator using statistical perspective. Applied Mathematics and Computation, 149(2004), 327-335.

S. R. Sain and D. W. Scott. Comment of "Flexible smoothing with B-splines and penalties". Statistical Science,11, 2(1996), 114-115.

R. Servien. Estimation de la fonction de répartition: revue bibliographique. Revue de Statistiques Appliquées, 150, 2(2009), 84-104.

J. C. Sevy. Convergence of iterated boolean sums of simultaneous approximants. Calcolo, 30(1993), 41-68.

J. C. Sevy. Lagrange and least-squares polynomials as limits of linear combinations of iterates of Bernstein and Durrmeyer polynomials. Journal of Approximation Theory, 80(1995), 267-271.

B. W. Silverman. Density estimation for Statistics and Data Analysis. Monographs on Statistics and Applied Probability, 26, Chapman & Hall, 1986.

M. A. Stephens. On the half-sample method for goodness-of-fit. Journal of the Royal Statistical Society, Series B, 40, 1(1978), 64-70.

R.A. Vitale. A Bernstein Polynomial Approach to Density Function Estimation. Statistical Inference and Related Topics, Vol. 2(1975), pp. 87-99.

G. Walz. Asymptotic expansions for multivariate polynomial approximation. Journal of Computational and Applied Mathematics, 122(2000), 317-328.

J. Wang and S. K. Ghosh. Shape restricted nonparametric regression with Bernstein polynomials. Computational Statistics and Data Analysis, 56(2012), 2729-2741.